A paper presented to the ESF-ORI First World Conference on
Research Integrity: Fostering Responsible Research
Held at the Calouste Gulbenkian Foundation, Lisbon, 16-19th September 2007

Empirical Realities of Scientific Misconduct in Publicly Funded Research

What can we learn from ORI investigations of U.S. cases in the biomedical and behavioral sciences?

By

Andrea Pozzi* and Paul A. David **

* *Stanford University:* pozzi@stanford.edu

** *University of Oxford and Stanford University*: pad@stanford.edu

PRELIMINARY DRAFT: DO NOT QUOTE WITHOUT PERMISSION

First draft: 26 June 2007
This version: 14 September 2007

ACKNOWLEDGEMENTS

The authors are grateful to the Science and Society Program of the European Commission (DG-Research), and Portugal's Ministry of Science, Technology and Higher Education for the support received for this research. We thank Dr. Nicholas Steneck of the Office of Research Integrity for his comments on the abstract and a preliminary set of slides based upon tables and figures prepared for this paper. The results and views expressed here are those of the authors and do not reflect official positions of the Government of Portugal, the ORI, or the European Science Foundation.

EXTENDED ABSTRACT

This paper presents the results of an exploratory statistical study of the nature and circumstances of instances of scientific misconduct among publicly funded researchers in the biomedical and behavioral sciences. Our analysis is based upon the of data made publicly available by the Office of Research Integrity within the U.S. Public Health Service for the period beginning in 1994 and ending in the first half of 2007, concerning its investigations of allegations reported to it by institutions that were recipients of federal research grant and contracts (pursuant to 42 Code of Federal Regulations §50.102).

Empirical studies such as this one are called for by the growing public attention and concern to the phenomenon of "scientific misconduct," which, in the discussion of this paper pertains to acts of plagiary, falsification and fabrication of reported research findings. Concern about these breaches of the "cognitive norms" upon which the integrity of the scientific research process depends certainly is well founded on a number of counts, whether or not the relative incidence of such acts on the part of researchers is rising -- a matter that has and will most likely remain very difficult to establish. In view of the growth of public and private sector expenditures, the continuing expansion of the volume and diversity of scientific research activities, and the central role of scientific information and capabilities in the economic, political and cultural aspects of modern life, there is a pressing need to inform public discussion and policy-making, both with in the research communities and the society at large about the circumstances and patterns that are manifested by this form of deviant individual behavior, and its relationship to the institutional and social contexts in which scientific research is being carried on. Lacking such knowledge, it would be difficult to begin to frame such preventive and corrective measures as may be warranted. Yet, that need has remained largely unaddressed. Although particular cases of intentional "scientific fraud" have been examined in detail, and these have in some instances been informative -- illuminating the psychological states and motives of the individuals, as well as the professional and occupational milieu in which particular breaches of the cognitive norms occurred and came to light -- there is remarkably little systematic empirical research to help place the insights drawn from those selected episodes in a broader context of statistical regularities. That has been due in good part to the paucity of systematic data collection, and to the inherent selectivity and biases in the gathering of information of suspected and proven instances of misconduct.

As a result of the work of the Office of Research Integrity, it has become possible to at least make a start towards putting in place some of the empirical foundations upon which intelligent and effective policy design must rest. At this stage we make no attempt to test statistical hypotheses about causal relationship, but examine the trends in the volume and distribution of types of misconduct that are reported, and of the situations of the alleged perpetrators within the respective occupational hierarchies of their research institutions. We utilize such data as is available to examine whether there have been temporal changes in the circumstances and basis for those allegations of misconduct (of different types) that were brought to the attention of appropriate institutional authorities. Differential propensities for the three types of alleged misconduct to be confirmed, and the way these have behaved over time, allowing for the duration of the investigation process, also can be examined. To investigate the influence of exposure to detection in the case of the different types of misconduct is possible to examine the relationship between the number of publications from the project prior to the reported instance of misconduct and the nature of the allegations, and of the confirmatory findings, respectively. Other findings relate to the association between the occupational status of alleged miscreant researchers and the (estimated) frequencies with which the misconduct allegations (of different types) brought against them are confirmed by the ORI's investigation. The paper discusses some possible interpretations

and implications of the empirical results, and concludes with some suggestions for collection and disclosure of data that would support more sophisticated and informative analyses.

1. Introduction: Background and Motivation for Research on Scientific Misconduct

The topic of scientific misconduct has been receiving an increasing amount of attention in the last twenty years, not only from an academic perspective (mainly in sociology and philosophy of science) but also in terms of government concern and mass media coverage. This might seem puzzling at first sight, seeing as scientific misconduct is not at all a new phenomenon. Zuckerman (1977) quotes different sources to report cases of presumed misconduct involving characters such as Ptolemy, Newton and Mendel[1]. What we have to consider, if we want to make sense of this surge of attention, is the extent to which, in the last century, the well being of the society as a whole has become to depend on the progress of science and, as a consequence, on the behavior of the members of the scientific community. While the link between scientific and technological progress and improvement in living standard is becoming more and more apparent, the scientific community is not able to support itself. This is not a new fact either, since science has always had patrons (David, 2004), for example in the form of kings or princes. Nowadays patronage by healthy individuals has become less relevant and the public operator has stepped in providing a large chunk of the funds needed to fuel research. As a consequence, the scientists find themselves not only carrying on an activity that has a huge impact on the welfare of the society at large, but also doing so as agents of the society itself, relying on public money.

Given that we are stressing the policy relevance of the phenomenon, it might look puzzling that we are presenting a fairly descriptive analysis. It is a long learned lesson that policy recommendations should not be based on descriptive models and we are by no means challenging this vision here. Still, we believe there are reasons that make worth even a descriptive study on the topic of scientific misconduct. A fairly obvious one is that there is remarkable scarcity of quantitative analyses in the field. Models of misconduct proposed by theorists can shed some light on the mechanics of the problem but ultimately deliver conclusions that are implicit in the (non testable) assumptions on the behavior of the agents. Reports from governmental agencies provide very useful statistical studies of the available data but do not make an extra effort trying to explain what they might suggest in relation to the deeper questions linked to the study of misconduct. In such a landscape, even a fairly descriptive analysis can unveil interesting patterns not evident just by looking at the aggregate of the tabulated data. Our approach starts from the same plain statistical analysis available in governmental reports and aims at pushing it to the limit, combining together in an original way different pieces of evidence to gain

[1] She acknowledges though, that scientific misconduct reflects a modern concern and that those examples are somehow anachronistic

information that would be overseen if we were to look at them separately. Even though the results we will present will be often nothing but suggestive, we believe that an assessment of what can be done with available data could serve several purposes. First of all, it will provide additional arguments to the debate over misconduct both in the academic and policy arena. Second, it might motivate other researchers, or the governmental agencies themselves, to look in greater depth and with more critical eye to the data already available. Lastly, to the amount of which we show that, with the data publicly at hand, it is not possible to give more than educated guesses with respect to many relevant questions, we hope to give impulse to the collection of more detailed data and to an increase of the sharing of such data between governmental agencies and independent researchers.

1.1 Social norms, social deviance and the special concept of "scientific misconduct"

In approaching the analysis of scientific misconduct, a first challenge is to provide a sharp operative definition of the phenomenon. Other than consensus on the fact that it involves the breach of some set of norms in force within the scientific community, there is no general agreement on what scientific misconduct is exactly. A useful taxonomy is provided by Zuckerman(1977) who highlights three sets of norms that can be violated. The violation of *cognitive norms* relates to undeliberate disregard of prescription of practices commonly believed to reduce the risk of generating invalid results (the so called *honest error*).

The violation of *moral norms* involves acts against rules that are of ethical significance to the member of the scientific community. The violation of *scientific etiquette* consists of deviance with respect of norms that govern personal relationships (i.e. eponymizing oneself, personal attacks to colleagues, and so forth). In this work I will only be concerned with the second category of violations. Violations of scientific etiquette are a concern since they are corrosive of the trust and reciprocation essential to the cooperative behavior that can greatly benefit scientific research. However, it is nearly impossible to figure out which ones of these practices do corrode trust and to what extent; hence I will ignore them. Violations of cognitive norms can harm even more seriously the work of the scientific community: undetected, undeliberate errors can mislead those who build new research upon flawed results. Nevertheless, since I finally aim towards a rational model of scientific misconduct, willingness and consciousness of the action and awareness of its consequences are assumed for each agent[2]: this is not the case for the honest error.

[2] This point can be controversial. If we allow for subjective beliefs, an agent can attach zero probability to outcomes deemed likely by other agents. From an empirical point of view, we would observe a behavior that we would be tempted to judge irrational. But preferences have no empirical content if not in terms of choices; hence, we can always think of a preference structure or of a system of beliefs that make these choices rational for the agent who makes them. For example, a scientist can have a strong taste for quick publication, even if they have to be retracted later on.

Although agreement on this selection is not general (Schmaus, 1983), this seems to be the orientation of the two main government agencies involved in research funding.
The National Science Foundation defines scientific misconduct as:

> *(1)fabrication, falsification, plagiarism or other serious deviation from accepted practices in proposing, carrying out, or reporting activities funded by NSF or (2) retaliation of any kind against a person who reported or provided information about suspected or alleged misconduct and who has not act in bad faith*[3].

while the definition of the Public Health Service (PHS) relates

> *...fabrication, falsification, plagiarism or other practices that seriously deviate from those that are commonly accepted within the scientific community for proposing, conducting or reporting research. It does not include honest error or honest differences in interpretation or judgment of data*[4].

This definition was review in 2004 and currently

> *Research misconduct means fabrication, falsification, or plagiarism In proposing, performing, or reviewing research, or in reporting Research results*[5].

We adopt this last definition, limiting our empirical analysis to cases of fabrication, falsification, and plagiarism. We will therefore leave out all those "*other practices*" to which both NSF and the old PHS regulation make reference. This does not mean that we are taking a stand in favor of the National Academy of Science in the quarrel (Buzzelli, 1993) between them and the government agencies about the best definition of scientific misconduct. We are aware of the relevance of several malpractices falling into the "other practices" category and quite convinced that they should remain included in a broader definition of misconduct. Circumstances falling into the "other practices" category, however, do not represent violations that are peculiar to the scientific community. Sexual harassment or funds distortion would not be any more tolerated in a law firm than in a research lab. Since our interest is not on deviance itself but on deviance taking place in the context of the critical relationship between scientific progress and social welfare, we will concentrate only on acts characteristic of this domain.
From here onwards, when talking of scientific misconduct we will refer to plagiarism, falsification, fabrication, or a combination of these three.

---

[3] 42 Code of Federal Regulations Sec. 689.1(a).
[4] 42 Code of Federal Regulations Part 50.102.
[5] 42 Code of Federal Regulations Part 93.103

1.2 Are anecdotes a basis for informed policy? A brief view of the literature

The bulk of the theoretic advances in research on misconduct lies in the area of sociology and spills from Merton`s work on the sociology of science. We already mentioned how part of this literature (Zuckerman, 1977; Schmaus, 1983; Buzzelli, 1993) was useful to setting the boundaries of what we will consider as misconduct for the sake of the present study.
In the same field, we can find several theories meant to explain scientific misconduct.
Hackett (1994) mentions three possible explanations: *individual psychopathology*, according to which scientific misconduct is due to mental disorder of specific individuals; *anomie*, which explains misconduct as arising from the tension between cultural goals and structural opportunities for research; and *alienation* that blames the overshadowing of the satisfaction from science by career strategy for the existence of misconduct. Zuckerman (1977) in her review lists the *labeling theory*, which is however oriented towards an explanation of repeated misconduct; *differential association theory* predicting that misconduct will be committed by those who associated themselves with former deviants; and the *conflict theory* that points to the way the group in power within the scientific community has the privilege to set the rules and decide what is deviant.

While this stream of literature provides a useful overview of the problem in broad terms, it does not provide a framework for formal modeling. A quite different approach to deviant behavior, useful to this purpose, is the economics of 'crime and punishment' (Becker, 1968). In the same line, Lacetera and Zirulia (2007) propose a game theory model that encompasses all the stages of a research projects, from the selection of a topic to the publication of the paper, and show how misconduct can arise with positive probability in any equilibrium of the game. The assumption of this class of models is that the agent is rationally comparing costs and benefits of committing a crime, taking also into account the probability of detection. He will engage in criminal activity if this is the action that maximizes his expected utility. This is, of course, a dramatic simplification, which we do not necessarily believe to represent the true process generating misconduct. However, if we think of misconduct as a planned act, it sounds natural to imagine that it must be triggered by the aim for some payoff and it will most likely involve a strategy meant to reduce the probability of being caught, which implies awareness of the risk of penalty. Whether those elements of rational calculation represent the backbone of the process or are of less important than other 'irrational' determinants, is matter of opinion.

Some work has been done in relation to specific types of misconduct: it is somehow remarkable that the closest attention has been devoted to plagiarism, defined by Rosamond (2002) “the most grievous academic crime”. Bartlett and Smallwood (2004) dig into four cases of plagiarism to raise concerns about the extent of the phenomenon and the little awareness of it in the public opinion. Ercegovac and Richardson (2004) also analyze plagiarism: their analysis is both a theoretical discussion of what is plagiarism and an attempt to understand what is driving what they believe is an upward trend in the phenomenon. They rely on figure published by Bronfenbrenner et al in the book “The State of the Americans”. Woessner (2004) builds a rational model of plagiarism, where

the agent is a student trying to plagiarize his term paper for a class. Plagiarism is seen as a gamble: it reduced the effort if undetected but might have costs if the student is caught. Woessner analyzes how the expected value of plagiarism varies under different sanctions against it (lowering the grade, failing the student, and so forth). For extensive survey of the literature on plagiarism, we refer to Enders and Hoover (2004 and 2006).

1.3 Modern concerns and government regulatory actions

1.3.1 The introduction of U.S. Federal regulatory actions

The first piece of legislation explicitly related to scientific misconduct is the Health Reserch Extention Act, voted by the Congress in 1985 in response to growing concern towards misconduct in publicly funded research in sciences. The Act recommended to give to awardee institutions guidelines to report scientific fraud and asked the NIH to be prepared to review those reports. The guidelines were produced under the form of the "NIH Guide for Grants and Contracts", published in July 1986. The final outcome of this legislative input was the well known "Responsabilities of Awardee and Applicant Institutions for Dealing With and Reporting Possible Misconduct in Science", that goes under the name 42 CFR Part50, Subpart A. This piece of regulations, effective since May 1989, represented the legal background of the activity tackling misconduct, giving precise definitions, dictating procedures, and prescribing actions.
The Part50, Subpart A has been the milestone in scientific misconduct regulation until very recently when the opportunity was taken to revise it, both to incorporate new instances arisen during the almost 20 years of activity and to update the federal regulations in accordance with new pieces of legislation such as the "Federal policies and Procedures on research misconduct", issued by the Office of Science and Technology Policy in December 2000. The new regulation, 42 CFR Part93 or "Public Health Service Policies on Research Misconduct", was published for comments in April 2004 and had its final letter in May 2005, replacing 42 CFR Part50 Subpart A on the 16th of June 2005.
The interest of the legislator in the issue of scientific misconduct is still alive, as other laws have been voted to help the governmental agencies in their action. An example is the Federal Whistleblower Protection Act (U.S.C. Par. 1201) meant to give additional guarantee to individuals who witness and denounce episodes of misconduct.

1.3.2 The creation and evolving mission of the PHS-ORI

As described above, the legislative action against research misconduct took impulse in the mid eighties; around the same time the condition for the creation of an institution such as the Office for Research Integrity started to create. Until 1986, allegations of potential misconduct in research funded by NIH grants were addressed to the granting institute. In 1986, the Internal Liaison Office took charge of receiving and addressing such complaints. This centralization effort went further in March 1989 when two different offices were created whose objective was to take responsibility of research misconduct: the Office for Scientific Integrity (OSI) and the Office for Scientific Integrity Review (OSIR). In May 1992 the two offices were funded to give birth the Office for Research

Integrity whose statue had already been voted by Congress (42 U.S.C. Par 289b). In June 1993, the NIH Revitalization act took responsibility over misconduct cases away from the granting institutes and gave it the ORI, established as an independent institution within the Department of Health and Human Services. ORI began publishing newsletters in 1993, while also starting its intramural research program. The first ORI annual report was published in 1994.

In 1999, ORI underwent a revision: it ceased to have directly part in investigations, role that played along with institutions in the first years of its life and its mission was stated to lie in the prevention of research misconduct and the promotion of research integrity. The responsibility for inquiries and investigations on allegations was left to the extramural institutions (the intramural projects in case of alleged misconduct in intramural research) and since then ORI only oversees and reviews procedures and findings of institutional inquiries and investigations. On the other hand, ORI took up a bigger task in the field of education starting massive training programs to promote responsible conduct in research as well as compliance with regulation on research misconduct. The ORI is nowadays fully active in dissemination and research activity on scientific misconduct through training programs, organizations of seminars and conferences, intramural and extramural grants.

1.4 An agenda for exploratory empirical studies in the biomedical and behavioral sciences

1.4.1 Motivation for focus on the phenomenon of scientific misconduct in NIH-funded research

While we are trying to say something in general about the phenomenon of scientific misconduct, our empirical evidence comes from a single source that spans the field of biomedical and behavioral science. We provide two reasons to justify this focus.
The first one is that said field is quite broad, displays remarkable breadth (from the study of mental illness and behavioral disturbs to the research on cancer therapy) and is prominent under many respects. Generally speaking, it is a field characterized by fast progress, or at least, continuous learning and, detail not negligible, blessed by some abundance of funding. Hence, focusing on it is of interest because on the success of this field depends to a large extent the future well being of mankind and because it is eating up a considerable amount of the pie of public funding to research.

A second reason, more factual, is that, for biomedical and behavioral sciences performed within NIH we had the chance to use data available through the ORI annual reports. It was important for us to start off our analysis without any need to get in touch with government agency, in order to preserve an external and impartial view. The availability of the ORI reports, with all their limits and the problems of a manual data collection, allowed us to build a dataset without any need to negotiate condition of use with the government agency.

1.4.2 Basic questions for quantitative analysis

Our analysis will touch several topics, relying on different pieces of data. In section 3.1, we time series on aggregated flows on allegations, inquiries and investigations to look at average rates of case disposition and frequency of findings, also conditioning to the type of charge investigated. In Section 3.2 and 3.3 we use the microdata from the case summaries of the ORI reports to investigate the distribution of time lags in misconduct correction. In Section 4, we exploit once again the microdata to look at effects of the available covariates on the conditional probabilities of findings of misconduct.

2 ORI Publications and the Sources of Systematic data

2.1 The Annual Reports and the ORI case management process

The data we use for this study come from the ORI annual reports, official documents published every year by the ORI that summarize the activity in the previous year. It provides information on the achievements of ORI in each of its field of competence: oversight of cases of scientific misconduct, dissemination and prevention, research activity, and institutional compliance. Our data come from the section "Responding to Research Misconduct Allegations" that provides aggregate information on the flow of allegations, inquiries and investigations under ORI oversight. Before looking at the structure of our data, it is helpful to explain what those data summarize, that is the management of research misconduct cases.

As can be seen in Chart 2.1, every year there is a universe of allegations of misconduct, that is, someone raises concern about the scientific integrity of a research study. The allegations can be made to the institution where the project is ongoing (e.g. a university), or to ORI directly[6].

If the whistleblower referred to the institution (blue arrow in the Chart), the institution will examine the allegation and decide either to dismiss it or to pursue it by opening an institutional inquiry. In either case, it is not compulsory for the institution to inform ORI; the process can go to the end of the inquiry phase without the institution being required to issue any formal report. However, if the institutional inquiry does not dismiss the allegations and it is decided to proceed with an institutional investigation, a report on the inquiry step is due to ORI that exercises his oversight function, checking procedural correctness and fairness of the inquiry. After the oversight of its inquiry, the institution can proceed with the institutional investigation; a report on the conclusion of this step is due to ORI for oversight and final assessment of whether the case will be closed with or without an actual finding of misconduct. In the former case, the report is forwarded to PHS that takes the administrative action. The administrative actions taken vary by case and include debarment, imposition of retraction of published articles, proibition from serving in any advisory capacity for PHS, and so forth. The respondent can appeal the

[6] It also possible that the allegation is made to NIH. For the sake of clarity we omitted this case from Chart 2.1. In any case, NIH would refer the allegation to ORI and, from there onwards, the allegation would follow exactly the same path as if it had been made to ORI to begin with.

decision and call for a hearing, challenging the outcome of the investigations; after the hearing the administrative actions become effective.

A slight difference in the path of case management is noted when the allegation is made directly to ORI (red arrow in the Chart). ORI initially reviews the allegation to check that it meets minimal requirements such as being related to research funded (or to application for funds) by PHS, that it meets the definition of misconduct set forth in the relevant regulation (42 CFR Part 50 Subpart A until June 2005, 42 CFR Part 93 after that) and that there is sufficient information to proceed with an inquiry. If the allegation is lacking under any of these profiles, the case is administratively closed. If the allegation is not dismissed, two things were possible until 1999. The first possibility was that ORI took charge of the inquiry and investigation step itself (solid red line in the Chart), all the way up to the declaration of a misconduct finding, or lack thereof. ORI lost the investigative prerogative when it was reformed in 1999 and nowadays only the other alternative (dashed red line in the Chart) is available. ORI will report the allegation to the appropriate institution; the institution will carry on the inquiry and, eventually, the investigation, leaving to ORI only an oversight role. Unlike the case where the allegation was made to the institution, however, the institution has to forward to ORI a report when closing the inquiry step, even if the outcome does not call for an investigation.

Chart 2.2 shows which part of this process gets to be reported in the ORI annual reports; in particular it refers to aggregated flows of misconduct, as opposed to individual microdata, the other data source underlying our study. First of all the chart shows that we have no information on the universe of allegations. In fact only the minority of the allegations is made to ORI and institutions do not have to report on the allegations they dismiss. Hence, we cannot know how many times the whistle is blown in a given year. We can however get an estimate by looking at the "Annual Report on Possible Research Misconduct" a form that all the institutions that want to be eligible for PHS funding have to submit every year with aggregate data on the number of allegations received and inquiries they are carrying on. The estimate on the total number of allegation from 1994 to 2005 shown in Chart 2.3 (3,571 allegations) is obtained summing up all the flows from the Annual Reports on Possible Research Misconduct and is, most definitely, a lower bound. We do have information on the exact number of allegations reported to ORI every year and how many of these are dismissed (red solid arrows in Chart 2.2).
For the inquiry phase, the ORI reports will list only those inquiries for which a report has been sent. This means that we can keep track of every inquiry that was triggered by an allegation made directly to ORI (in that case the institution has to issue a report) and of all the inquiries born by an allegation made to the institution that called for an investigation. In other word, the total number of allegations pursued displayed in Chart 2.3 is an underestimate since it misses the inquiries generated by allegations made to the institution and that did not call for an investigation. Whether the underestimate of the number of inquiries is larger of smaller than the one in the total number of investigation that we were mentioning before it is hard to guess and impossible to know.
Finally, Chart 2.2 shows how we have full information on aggregate flows from the investigation step onwards: reports are due to ORI on every institutional investigation and

it is kept record of the outcome of the investigation, whether it closes with or without findings of misconduct.

It is important to remark how the ORI revision in 1999 changed the shape of the process: from 2000 onwards, no inquiry or investigation was taken up by ORI directly. All the allegations were referred to institutions for institutional inquiries or investigations. For the sake of simplicity, in the paper we use sometimes expressions as "ORI investigations" or "ORI's investigations". By that we do not mean that the actual investigation was performed by ORI since we know it cannot be the case after 2000 and that even before it was a fairly rare occurrence[7].

The figures we rely upon for the aggregate flow analysis presented in Section 3 of this research were manually extracted from summary tables in ORI annual reports 1994-2006 to create a 13 years time series for several series (number of allegations, number of inquiries, number of investigations, outcome of investigations, and so forth).
The main limits of the ORI aggregated flow data are the same as any statistic dealing with some kind of illegal activity. First of all, we only get information on cases where a complaint was made: this might be the "tip of the iceberg", since we can imagine that in many instances misconduct can be performed without raising any suspicion. Second, even if there are informal concerns about the integrity of a research project, no track is left in the reports if no one decides to blow the whistle. It is possible that some share of the misconduct performed it is detected but no action is taken against it either because no one dares to report it (see popular press for accounts of potential troubles faced by whistleblowers) or because the institution decides to keep it quiet. This might be especially the case when allegations are made against very prominent scientists that are both key to the reputation of the institution and more capable of influence its decision and prevent the allegation to ever be reported to a government agency. Of other limits, relating to problems/mistakes in reporting and information that could but it is not made available; we will talk in later sections.

Chart 2.3 makes clear that the amount of cases that gets under deep scrutiny, that is to the investigation stage, is a very small fraction of the original number of allegations. While there is nothing pathological in it (it is perfectly plausible that most of the allegations are not relevant), this limits the size of our sample with regard to some of the most interesting exercise we carry on. In fact, only for cases reaching the investigation phase the information contained in the ORI reports becomes rich enough to allow for more than a mere tabulation of numbers. First of all, as shown in Chart 2.4, only for investigations we have a reliable breakdown of flows by charge. As it will become clear, there are several reasons to be interested in looking not at generic allegations of misconduct but at pattern for specific charges. For example, to the purpose of policy, we might be more concerned with fabrication and falsification than with plagiarism that is, from the point of view of the collectivity, a "victimless crime". Careful exploration of this dimension is only

[7] ORI closed 2 investigations (plus a comparticipation in other 7) out of26 closed in 1994 (27%), 9 out of 41 in 1995 (22%), 4 out of 38 in 1996 (10.5%), 2 out 29 in 1997 (7%), 3 out of 21 in 1998 (14%), and 0 out of 25 in 1999 (0%).

possible for cases that reached the stage of investigation, since information on the type of allegations is only provided for investigations for the 1994-2006 period[8].

Just as it happens with the total flow of investigation, we could think of tracking a single misconduct case from its origin (Chart 2.5). The path is the same explained for the total flow of cases but Chart 2.6 shows very clearly that we know a lot less about individual cases than we do about aggregate numbers. In fact, investigations are the only source we can use to build microdata that allow a more careful analysis of misconduct cases. A short case summary of closed investigation is published as an appendix to ORI annual reports. For investigations closed with findings of misconduct name of the respondent, degree, position, affiliation, granting institute(s), charge(s) and actions are made available; gender is not explicitly stated but can be inferred. In the case of investigations closed without conclusive evidence, confidentiality and privacy issues force the report to be far less informative: only the position of the accused, the alleged charges and the granting institute(s) are mentioned. Manually coding information from the case summaries of the ORI annual reports, we built a dataset that we exploit for the microdata analysis presented in section 4. We discarded case summaries (very few) that related to the so-called "other charges" and ended up with 351 observations, 158 investigations closed with findings and 193 investigations closed without findings. The benefit of building such a dataset is, in our opinion, significant. It allows looking not only at conditional distributions as in the case of aggregated flow data (e.g. distribution of findings by position) but also for multiple controls at the same time (position, granting institute, etc…) a great improvement towards detection of significant correlations that are not spurious. Of course the microdata are also "soft data" to some extent. First of all, the different richness in information between investigations closed with and without findings limits what we can do with it. For example, we cannot use the degree of the respondent or his gender as regressors in the microdata analysis since we know that information for cases closed with findings but not for cases closed without findings. Other information we would like to have is altogether missing, such as a mention of the identity of the whistleblower. The fact that we had to code the data manually represents an additional weakness. The verbal description is not always clear enough to extract all the information needed (e.g. many cases in which we are unable to state the position held by the respondent) or leaves ambiguity (are "research technician" and "laboratory technician" meant to identify the same role or two different ones?).

Despite all the problem that can be devised with the data collected from the ORI annual reports, we believe that our effort in coding aggregate flow data and, especially, microdata from the case summaries produced an amount of evidence very useful to push the limit of inference far beyond what was done before with the same data.

2.2 Previous statistical research on ORI data

2.2.1 Internal and extramural research

[8] Breakdown by charge for inquiries is reported in the ORI reports only since 2000.

ORI carries on a very broad research activity on themes linked to misconduct, which also implies effort to gather very diverse datasets as basis for the analysis. The intramural research misconduct has touched, in the last few years, topics as the reporting of suspected misconduct, the trends in closed investigations, the creation of an environment favorable to research integrity, and an assessment of the effectiveness of its own education program[9]. In the meanwhile, the beginning of Research on Research Integrity (RRI) Program has fostered extramural research on misconduct, especially in relation to the factors affecting integrity in research. The extramural research program started in 2001 and, by the end of 2006, had already helped generating more than 30 journal publications.

This remarkable research activity notwithstanding, a paucity of previous work on the determinant of misconduct from the point of view of the information provided from the investigative action remains. Out of the 17 research completed in the intramural program, very few analyses data from actual inquiries and investigations, while many more use surveys or case studies. As we already stated, we believe it is key to keep looking at evidence from actual misconduct cases since, even though they are subject to selection issues. If the objective is to educate and prevent research misconduct, it is clear that some lesson has to be learn from data that are indeed related to true cases of misconduct; relying entirely on anecdotal evidence or survey (where people report on what *they* think it is misconduct) it is not advisable.

The data on aggregate flows we use in this study has indeed been study by ORI internal researchers; here we want to single out the statistical review of these data made by Rhoades (2004) to which we are indebted in two ways[10]. First of all, it has represented a source of extra data to us: Rhoades tabulates the distribution of inquiry and investigation data with respect to many dimensions no longer present in ORI annual reports[11]. Rhoades provides distribution of data conditional on gender, position and education of the respondent, granting institute and funding mechanism, and gender, position and education of the whistleblower. Without his very accurate study, we would have not had access to this information. A second important feature of Rhoades' contribution is that it presents many research questions, and very relevant ones, that should be addressed based on available data. He limits to suggest directions for future research and does not venture to go beyond a statistical display of data; we tried to take up some of his challenges and, with the help of the same data he uses and the extra material coming from our microdata, to give answers to some of the question he poses.

---

[9] The mentioned projects refer to "Reporting Suspected Research Miscnduct in the Biomedical and Behavioral Research" (2006), "ORI Closed Investigations into Miscondcut Allegations Involving Research Supported by the Public Health Service: 1994-2003" (2004), "Integrity in Scientific Research: Creating an Environment That Promotes Responsible Conduct" (2002), and "ORI Education Program: A Needs Assessment" (2001). A full list of ORI intramural research is available, with possibility of downloading the documents, at ORI's website, http://ori.dhhs.gov/research/intra/studies_completed.shtml, last accessed on 09/14/2007.

[10] An earlier study, published without author and very similar in the spirit to Rhoades' one is "Scientific Misconduct Investigations, 1993-1997" (1998).

[11] The structure of Rhoades research follows closely, in the order of the tables, the one of the "Statistical appendix " to the ORI annual report, published between 1994 and 1997.

In relation to Rhoades' study, it has been brought to our attention that, comparing closely the figures; some disagreements in the number arise. This should not be the case, since we reference the same source, and triggered us to a deep check of our data to reconciliate them with Rhoades' ones. Three reasons seem to explain the divergences between the two studies. First, there are a small number of factual mistakes in the statistics displayed in ORI reports. For example it happens that a group of allegations is counted twice in the official report; Rhoades accept the total presented there, while we corrected what appears to be a misprint and get, therefore, a lower total number of allegations for that year. A second source of disagreement is due to Rhoades using, presumably, revised data with respect to the ones published in the ORI reports. For some of the numbers in disagreement with Rhoades (2004), we check our original source, the ORI reports, and found that our coding was accurate; Rhoades however departs from the number published there. Since those are small adjustments, we guessed that Rhoades compiled his study relying on revised statistics, where the year of competence of a given case might shift, altering the total of cases in two contiguous years. Finally, sometimes our and Rhoades' studies disagree on counts because we made different choices or had different focus. We decided to disregard the "other practices" of misconduct, while Rhoades keeps them in his totals, or, for example, to count cases where fabrication and falsification were performed at the same time as increasing the total count of both fabrication and falsification cases, while Rhoades counts them in a separate category. For a more detailed analysis and reconciliation of the divergences between our and Rhoades' annual totals, we refer to Appendix A.

3 Dynamics of the Phenomenon: The view from ORI, 1994-2006

3.1 Absolute and relative trends: is the problem exploding?

In this section, we start exploiting the aggregated flow data to perform a time series analysis of trends in scientific misconduct in biomedical and human sciences. The analysis presented two main methodological challenges. The first one is due to the fact that our data span starts in 1994, when ORI had just begun its operations and concludes in 2006. We are aware that ORI did not start its activity abruptly in 1993 but was preceded by OSI and OSIR and that misconduct cases were handled even before within NIH. However, it is clear that the early 90's were a period of big revolutions in the field of the fight to misconduct; as a consequence, we believe, the initial conditions (the case load inherited from the past, to mention one) might have played a role in the way ORI acted in the first two years of its existence. An indication of this is given in the Annual Reports for 1994 and 1995, where it is stressed how the case load has been reduced, that sounds like an implicit allusion to a "normalization" goal. At the same time, our data collection reaches 2006, while the operations were still going on and this is likely to introduce some distortion in our numbers. For example, if investigations closed with findings take longer than investigations closed without findings, at the end of our sample we will have a prevalence of investigations closed without findings only because they got in the records

quicker than investigations of the same cohort that are about to be closed with a conviction. We deal with these issues with several techniques to smooth data: either we average the first two years of operations or take two years moving averages of the whole series.

A second concern is due to the fact that fluctuations in the number of the cases opened every years and the length of the process makes it difficult to compute meaningful statistics for the case management. For example, what is the relevant benchmark for the number of investigations closed in a given year? The number of investigations closed in the previous one? But, if the number of open cases is different in the two years, the comparison would not be meaningful. Choosing the ratio between closed and opened investigation would also be problematic because it would disregard that some years have a higher number of open cases because of the ongoing procedures inherited by previous years. Our choice, then, is to consider as the relevant denominator for all the ratios related to case management activity the number of active cases including the ones opened in the year as well as those inherited from previous operations.

A look at Graph 3.1 reveals that both the number of misconduct allegations and the number of cases, out of the initial allegations, that reached the final stage of the investigative process were pretty stable in 1994-2006. The number of allegations displays a weak upward trend while the number of investigations closed seems to go downwards. Given the process of reform undergone by ORI in 1999, we perform a careful check of several statistics exemplificative of the case management behavior, looking at significant differences pre and post the shift of focus of the Office. Breaking the sample in this way, we think we are allowing the change in focus of ORI to display its full impact and that we are controlling for any possible effect of that shift. The full battery of tests is presented in Table 3.1 and the bottom line from it is that there is one single significant change. The passage between inquiry and investigation phase and even the conviction rates are remarkably constant and appear to have not been affected by change in reporting regulation, in definition of misconduct and in the scope of ORI investigative activity. All in all, the tests in Table 3.1 show no dramatic change after the shift in ORI's mission in 1999. This result is not unexpected: the main effect of the revision was to prevent ORI from actively taking part in inquiries and investigations and we already documented that this participation had always been limited.

What changed significantly is the ratio of the number of cases opened to the original number of allegations received, which shifted down after 1999. ORI investigates a lower fraction of the allegations of misconduct; this increased selectivity can have two explanations, which do not exclude each other. A first possibility is that there has been an increase in the number of blatantly unfounded allegations: the diffusion of internet and other ICTs has reduced the cost of checking on someone else's research, increasing the number of projects each individual is informed upon but, possibly, reducing the depth of his knowledge about it. This story would be consistent with a raise in unsubstantiated accusations. On the other hand, we can think that the proportion of founded accusation stays the same but ORI is unable to meet the increased demand for oversight and initial scrutiny. As a result, more allegations would be dismissed to keep constant the number of

cases examined. The scatterplot in Graph 3.2 points to the same stylized fact: in years where the number of allegations is higher, the number of cases opened does not rise in proportion. The same patter is obviously devised for the fraction of findings over the total number of allegations (Graph 3.3): if the fraction of cases opened reduces, opportunities for findings do as well. If the first potential explanation of the phenomenon is the true one, then ORI dissemination and educational effort will solve the problem in the long run, teaching scientists how to self select allegations worth making. If the mechanism behind were the second one, it would cast shadows over plans to shift up the scale of prosecution towards misconduct, since it would mean that ORI is already operating close to its capacity constraint.

Graph 3.4 and 3.5 are meant to point out that, despite the shift in selectivity we discussed above, the phenomenon of misconduct displayed great stability in 1994-2006. First of all, the number of allegations is increasing but not exploding. Moreover, if we trust the decision on the dismission of allegations, the problem is not growing at all, since the number of cases is oscillating by year but substantially stable. Ratios of inquiries and investigations closed by year and, most remarkably, their cumulated levels tend to be flat in the period of our analysis. Misconduct seems to be in a stationary state and this justifies our approach of pooling microdata from different year and using it as a panel in Section 4. From a policy point of view, it suggests that a more aggressive approach to avoid an explosion of the problem is not called for from data that rather display great stability.

3.2 The volume and the distribution of misconduct by type of allegation: The “morphology of misconduct”

If research misconduct looks like a stationary process, when looking at the aggregate number of cases, does this impression changes when we take a closer look? To this purpose, we analyze the breakdown of cases by type of charge, that is, distinguishing cases on the basis of whether the allegation was of plagiarism, of falsification or of fabrication. In the letter of 42 CFR Part 93.103, plagiarism is “the appropriation of another person‘s ideas, processes, results or words without giving appropriate credit”, fabrication is “making up data or results and recording or reporting them”, falsification implies “manipulating research materials, equipment, or processes, or changing or omitting data or results such that research is not accurately represented in research records ”.

After a downward trend in the first three years of operations, the number of investigations closed for each of the three types of sin looks stable (Graph 3.6). Moreover, the series for the three charges move together (particularly falsification and fabrication), suggesting that even more stable than the number of investigations is the share of the total held by each type of allegation. This suspicion is confirmed by looking at the top panel in Graph 3.7: the pattern is pretty consistent and shows a share of about 60% to falsification investigations, 30% to fabrication and 10% to plagiarism. The relationship between investigations and findings by type of allegations can be inferred comparing the two panels of Graph 3.7: there is no sin that imply a greater probability of convictions since

the share of the total number of investigations by type are very close to those of the total number of actual findings of misconduct by type.

A different perspective to look at the morphology of misconduct is to observe how respondents are distributed by type of allegation. Table 3.2 displays conditional distribution of investigations closed without findings while Table 3.3 present the one for investigations closed with conviction. It looks clear that there is "segregation" in the performing of misconduct and that, even conditioning for the type of allegations, the distribution varies when we consider whether the allegation was substantiated or not. Plagiarism seems to regard mostly associate and assistant professors both in terms of allegations substantiated and not substantiated. Falsification looks like a high rank academic affaire if we look at the allegations non substantiated but are mostly low rank academics and non academics that are convicted for such a crime. A similar pattern can be devised in the case of investigations with allegations of falsification and plagiarism jointly. Several lessons can be taken from the analysis under this dimension. First, different people perform different types of misconduct; this suggests that the reasons and the incentives for doing so can vary wildly for individuals at different positions in the field of sciences. This raises some skepticism on the notion that a single paradigm to tackle misconduct can be adopted. The optimal recipe should take into account that misconduct cases arise sometimes from desire of quick fame by high rank academics and sometimes from boredom of low rank employees or students doing alienating jobs in the lab. Methods to prevent those two types from deviance have to take into account the difference in terms of their goals, risk aversion and rationality.
A second striking finding is that there are dissimilarities between the distributions of investigations with and without findings, for most of the type of charges. In other words, high rank scientists are convicted much less than they are accused. This could be explained by an excess of attention towards the work of prominent researcher, that generates many undeserved allegations. Or with the ability of these "stars" to avoid conviction more than lower rank scientists can possibly manage. The recommendations for policy are completely different according to which one of the two story is the true one and we will return to the issue of the heterogeneity in the rate of conviction in Section 4, with the help of our microdata.

3.3 Some dynamic aspects of detection and investigation of scientific misconduct

After exploring some trends in the way misconduct is performed, we now move to examine the other side of the issue: the characteristics of the detection process. Two dimensions of this action are particular relevant in our opinion: the effectiveness of the investigative process in timely spotting research misconduct and the cost at which the detection activity is performed. The following two subsections are devoted to present our findings on these topics.
The importance of having estimates of the time lags with which misconduct is detected is linked to the social cost of undetected misconduct. We have previously shown that misconduct in biomedical and behavioral sciences do not seem to be exploding; if the lag at which misbehavior is detected is large enough, however, the fact that the proportion of

misconduct over time is stable is not enough to claim that the social cost is not rising. In fact, if research moves faster and funds are accordingly allocated faster, the distortive effect of undetected misconduct is increasingly relevant even if both the incidence and the detection lag are constant over time. Of course this is true if the magnitude of the detection lag is big enough with respect to the rate of progress of research. Therefore, an assessment of the actual length of the detection gap it is needed to realize the potential social cost of undetected misconduct.
To address the question of the cost of the detection process, we look at the length of the investigative steps to be taken in order to assess an allegation of misconduct. The length of inquiries and investigations it is linked to the cost of misconduct detection in two ways: directly, because investigating misconduct cases it is expensive and this activity diverts funds from other uses. For example, at the level of institutional inquiries and investigations, researchers are required seat in the investigative panel: the salary paid during the time they spend in the panel has an opportunity cost in term of the research they could have produced. But the length of the investigative procedures is costly also indirectly since it contributes to the correction lag, the time between the diffusion of a result tainted by misconduct and its retraction, which, we already argued, is costly to society.

Before proceeding to showing the actual procedure we followed and the results, it is worth making a small point about retraction, a key piece of information in the computation of our "correction lag". In the microdata, we had a total of 65 cases closed with findings that involved at least a journal publication, hence there was room for asking 65 retractions. We count 25 voluntary retractions, that is retractions sent by the respondent and published even before the case was formally closed; this leaves 40 requests of retraction potentially made by ORI. If we rely on estimate from a study by Neale et al. (2007)[12], we can assume that only 48% of these 50 papers will be actually retracted[13]. This would lead to 24 retractions, less than those voluntarily submitted by the authors. This suggests that incentivating voluntary retraction could be a way to have a higher number of corrections (and more timely ones) since imposition of retraction appears to be weakly effective.

3.3.1 Measuring the detection and the correction lags from the ORI investigations

Our approach to estimating the detection lag, the time between the disclosure of a research resulted from misconduct and its being flagged as such, is based on the collection of the microdata from the case summaries for investigations closed with findings of ORI annual reports. In order to calculate the detection lag, we would need the date at which the misconduct act has been performed and the date at which the allegations has been made; we have neither. We can, however, have an upped bound for the date at which misconduct occurred, at least for the subset of cases where a publication was

[12] We are thankful to Nicholas Stenek for pointing us to this paper. Other contributions in the recently increasing literature on retractions are those by Harold and Drummond (2006) and Atlas (2004).
[13] They analyze a sample of 110 cases where, out of 98 requests of retraction, only 47 papers were actually retracted. However, other forms of correction (e.g. publication of a comment or of an erratum) were put into place in some of the cases where no retraction was made.

involved. If the tainted research originated a publication, this will be reported in the case summary of the ORI report (since one of the action upon conviction is to ask for a retraction): hence we can assume that the misconduct episode occurred before the year of the first publication based on the tampered research. On the other hand, the publication of a case in the ORI report means automatically its exposition as an instance of misconduct. Hence, detection must have occurred before the year in which the case is mentioned in the ORI report: this is an upper bound of the detection date. Moreover, since we a re working only with cases where a publication is involved, we have a chance to refine this upper bound. As said, when a research project is assessed to be a misconduct case, the respondent is asked to send a retraction to all journals where outcome of that research has been published. The respondent can decide to comply with this request even before the investigation process is officially over: this generates cases where we have a retraction published in a year prior to the one in which the finding appears in the ORI report. In this case, we will take the year in which the retraction has been published as estimate of the detection time. To summarize (see also Chart 3.1), our estimate of the time at which misconduct was performed is always the year of publication of the first research linked to the case; while our estimate of the time of detection is the minimum between the year in which the case appeared in the ORI report and the year of the first retraction for the publications involved. Since our estimate of this time interval relies on the procedure of formal correction of a scientific finding based of misconduct, we call it “correction lag”.

The detection lag can be shorter or longer than our correction lag; what is sure is that the correction lag overestimates the time between the publication and the first allegation of misconduct. In fact, by construction, it includes the time of the inquiry and investigation process, which are performed after someone blows the whistle. Graph 3.7 present the distribution of the correction lag computed on the bases of the 65 cases of misconduct involving publications in our microdata; Table 3.4 reports relevant statistics on this distribution. The mean correction lag is 3.65 years and the median is 3 years, the correction lag ranges between 0 and 16 years. Graph 3.8 and Table 3.5 repeats the analysis considering only cases of falsification, which, it appears, take slightly longer to be corrected; Graph 3.9 and Table 3.6 present again result for the universe of cases but trimming our sample of an outsider in 1994 which affects the evolution of the distribution.

As said, 3.65 years is an overestimate of the time between publication and detection, since it includes the time spent in institutional inquiries and investigations. If we were able to remove this additional component we would find not the detection lag itself (for that we would need to know the exact date at which misconduct was performed) but we would surely have a better approximation of it. Rhoades (2004) provides us with an opportunity to do so.
Rhoades tabulates distribution of the caseload by length of inquiry and investigation steps, distinguishing on the basis of the outcome of the process. From these tables, we were able to infer the length of an inquiry that led to an investigation (68.07 days) and the length of an investigation closed with findings (168.72 days). Hence, the length of the inquiry and investigation process for a case closed with findings of misconduct is 236.79

days or 0.65 years[14]. Subtracting this to the average correction lag gives a refined estimate of 3 years for the detection of misconduct. How good of an approximation it is, is surely matter of opinion; we content ourselves, for the moment, of having being able to provide a number, no matter how rough, to start a debate on this issue. Whether 3 years is too a long time or a reasonably short one is not clear and cannot be decided without taking into the picture additional information as the rate of progress in science (which is, arguably, heterogeneous across fields) and the level of cumulativeness. If science progresses very fast through jumps that make previous knowledge obsolete, results originated from misconduct will soon become irrelevant – no matter how quickly or slowly they are detected- along with all the honest errors and the good ideas that would have deserved further exploration.

3.3.2 The duration of the investigation process and the entailed social costs

In this section we propose a heuristic to estimate the per case resource direct cost of the process of inquiry and investigation over misconduct cases; for the detail of the procedure, we refer to Section 2 in Appendix B. The particular number we will mention depend crucially on the assumption made all throughout the "guesstimation" process and have to be considered with many caveats. However, our aim has been to show that even with data already available, at first sight unfit to recover economic cost of the misconduct prosecution machine, some steps can be made. If this will stir some debate and prompt more attention and additional data collection, allowing substituting our numbers with more reliable ones, our effort will have been worth.

We start off making an assumption on the burden to sit in a inquiry or investigation panel for a member. We assume that a panelist devotes six hours of his time each month for every month the case is under scrutiny. Then, we rely on Rhoades' tabulations for the period 1994-2003. They allow us to obtain the average size of the panel for the 165 investigations closed with findings (3.65 panelists), the 102 investigations closed without findings (3.78) panelists and the 70 inquiries not leading to investigations (2.96 panelists). Rhoades data have already been exploited (see Section 3.3.1 and Section 1 in Appendix B) to obtain the average length of inquiries and investigations: inquiries and investigations closed with findings last, on average, 7.9 months; inquiries and investigations closed without findings last 8.6 months and inquiries that did not require investigations last 2 months.
This information is sufficient to compute the total number of hours allocated to examining misconduct: it amounts, for the period 1994-2003, to 28,472 hours for investigations closed with findings, to 19,972 hours for investigations closed without findings and to 2,483 hours for inquiries not leading to investigations. We further assume that the working year of an academic consists of 9 months, or 184 8-hours working days. This means that all the cases that reached the stage of investigation (those closed with and without findings) used up 33.26 salaried academic work years in 1994-2003. Inquiries not followed by investigation required 1.7 salaried academic work years. In total, the 337 cases examined in 1994-2003 costed almost 35 salaried academic work years.

[14] For a complete analysis of the length of the inquiry and investigation step, we refer to Section 1 in Appendix B.

A detailed discussion of the limit of this estimate can be found in the appendix, here it will suffice to say that, without further information, we cannot even tell whether our numbers are likely to be too big or too small. This should be a clear sign that extra effort has to be made to collect data suitable for more accurate estimate. Once we have reliable figures on the direct social cost of the investigation process, it will come the time to start debating on whether or not this cost is excessive.

4 Investigations of Misconduct and their Outcomes

In this section, we move to a descriptive microdata analysis exploiting information coded from the case summaries of the ORI annual reports from 1994 to 2007[15]. Even though data come in a time series fashion and nearly none of the respondents appears more than once, we choose to pool observations from different misconduct cohorts and treat them as if they came from the same population. One of the goals of the time series analysis we performed in Section 3.1 and 3.2 was indeed to show that there is enough stability in the phenomenon to justify this choice. In particular, it will be recalled that pattern of the investigative phase were shown to be remarkably stable throughout our sample period. Hence we proceed to pool observations from the 14 years time span and obtain a sample of 351 investigations, 158 closed with findings of misconduct and 193 closed without findings. As already mention, several limitation of the data will affect the extent to which we can push our analysis. It is worth recalling that: i) the information available for investigations closed without findings is far less rich than the one contained in reports for investigations closed with findings. This, for example, will preclude the possibility of controlling for gender of the respondent that is known in the latter case but omitted in the former. ii) Since the data have been manually coded there are data missing problems. It was not always possible to infer all the information needed from the case summaries; for example we were unable to retrieve the position held by the respondent for more than 30 of our observations.

The main objective of the microdata analysis is to recover the marginal effects of some characteristic of the misconduct case on the probability of the case to be closed with findings. Of course, since the sample only contains investigations, all the results will be conditional on the case reaching the stage of the investigations. Estimating the impact of similar covariates on the probability of an allegation to be selected for inquiry and for an inquiry to go on to the investigation stage would be most interesting but it was impossible given the lack of microdata on cases at the allegation or inquiry stage. We concentrate on a small set of explanatory variables for which we have information for both investigations closed with and without findings: position of the respondent, granting institute, involvement of journal publications in the case. Before moving to the analysis of the marginal effect and to the estimation of probabilities of conviction, we display, as a mean of presenting the microdata, some result from a univariate analysis of our sample.

[15] Even though the ORI annual report for 2007 is not yet published, case summaries for investigations closed with findings of misconduct are available on ORI's website. Our data are updated to June 2007.

4.1 The characteristic of the case: Position and granting institute

The case summaries in the ORI report mention 27 different positions for respondents in misconduct investigations. For the sake of our analysis, we decided to group them in a coarser partition as displayed in Table 4.1. In the table it can be noticed a double distinction: between academic positions (on the left) and non academic ones (on the right) and between high rank positions (on the top) and low rank (at the bottom). Our aggregation of positions in broader categories results in a quite balanced sample, where all the four typology (high and low rank academics and high and low rank non academics) are fairly represented. A caveat on the aggregation procedure is, of course, that if we arbitrarily aggregated in the same categories position that have actually heterogeneous effects on the probability of conviction, the coefficient obtained in the probit regression will be meaningless.

Graph 4.1 provides a quick look at the distribution of the conviction rate across positions. On the x axis, we have the total number of investigations in which the respondent belonged to a given category, on the y axis we have the number of investigations closed with findings with respondent of a given category. The ratio provides a conviction rate: a ratio of 1 means that members of the category are always found guilty if they reach the investigation stage; a ratio of 0 would imply that they are always acquitted upon investigation. If we pool cases from all types of allegation, we can identify three clusters: a high conviction rate group (found guilty almost 70% of the time they are involved in an investigation) that includes low rank academics: graduate students and post-docs. An intermediate conviction rate group (about 50%) that includes high rank non academics and a low conviction rate group (slightly above 30%) with high rank academics. Low rank non academics (the group labeled “staff”) are between the high and the intermediate cluster, while research assistant and undergraduate student group do not have a large enough sample size to allow for a comment.
Graphs 4.2 to 4.4 examine the effect of the type of charge on such a pattern. It is worth reminding that, in case of multiple allegations, we consider a case as belonging to as many typologies of charges as it has been made. For example, a case where there are suspicions of falsification and fabrication is for us a case of both falsification and fabrication. The result for the pooled charges case seems to be mostly driven by falsification, the charge represented most prominently in our sample, but there are other points of interest. In falsification investigation we observe an outlier (the post doc fellows) with a particularly large conviction rate and a group including other low rank academics and high rank academics with an intermediate rate. High rank academic display a much lower rate of conviction. Plagiarism seems to concern a smaller number of categories, consistently with our previous observations on the distribution of position across types of charges. Only associate and assistant professors, research scientists and post docs are significantly involved in this type of investigation, with the latter group being more convicted than the others. The graph for fabrication portrays a situation where we have heterogeneity in the level at which different positions are involved with investigations on this charge but substantial homogeneity on the conviction rates.

Another piece of information relevant to our analysis is the identity of the institute within NIH that was funding the grant awarded to a scientist later on involved in an investigation. Table 4.2 lists all the granting institutions present in our sample; for our analysis, we decided to single out the most prominent ones (where the threshold for prominence was arbitrarily fixed at 17 or more observations) and group the others. A large share of our observations comes from grant awarded from the National Cancer Institute and the National Heart, Lung, and Blood Institute but behavioral sciences are also represented through the National Institute for Mental Health. If we repeated for the granting institute the same exercise we made for the position held by the respondent, the result is quite different. Graph 4.5 shows a much smaller dispersion of the rate of conviction across institute than it was the case for different positions. In this picture of homogeneity, the top two represented granting institute show low than the average rate of conviction but the third and the fourth in the ranking have a higher than the average rate of conviction. In conclusion, while the analysis of the effect of the position opens many questions, the granting institute appears as a fixed effect to be removed from the analysis but leaves no puzzle to be solved.

4.2 Probit regression Models and Predicted Probabilities of Findings

The main purposes of the microdata collection has been to allow for fitting a binary outcome model to the data and enable to identify marginal effects of the regressors of interest.

In the simplest specification, the dependent variable is the outcome of the investigation, a dummy variable assuming value of 1 if the investigation was closed with findings and 0 if it was closed without findings. The explanatory variables included a set of dummies identifying the position held by the respondent among the nine categories in bold font in Table 4.1 and a dummy indicating whether a journal publication was involved in the case. The results from this probit specification are reported in Table 4.4. The model has been fit separately for different types of charge and the omitted variable in the set of position dummies is “full professor”, hence the marginal effects for the positions are to be interpreted with reference to the baseline level of the full professors. All the position coefficients are positive for the different charges, implying that the probability of conviction rises for positions other than full professors, not all of the position dummies are significant, though. Having a publication involved in the case increases the probability that the investigation concludes with findings of misconduct; this effect is significant for falsification cases. The interpretation of the “journal publication” dummy is not trivial: we would certainly expect it to have an effect on the probability of a misconduct case to receive an allegation, since publication increases diffusion of the results of a research and widens scrutiny over it. But once we condition on the fact of having reached the investigation stage, there is no immediate reason to think that publication should affect the outcome of the investigation.

From the result of the probit, we derived fitted probabilities of conviction that are reported in Table 4.5. The positions of the respondents are ordered in decreasing probability of conviction: a post doctoral fellow involved in an investigation for suspected falsification has a 50% probability to be found guilty, while a full professor in

the same situation will walk out free in 80% of the cases. The ranking changes for different allegations: graduate students are the most likely to be convicted for fabrication and assistant professors for plagiarism. In general, low rank scientists have much higher probabilities of conviction than high rank ones and the probabilities rise for all categories once a publication is involved.

To check robustness of these findings, we run a new, richer specification adding controls related to the granting institute. In this specification we control for all the factors already included and add a set of dummies denoting the institute(s) that was financing the research project; we also put dummies to pick up eventual effects linked to projects financed by multiple institutes at the same time. In table 4.6 we propose results of this specification only for cases involving falsification, the most prominent sin in our sample. The omitted position category is still “full professor” and the omitted granting institute is the National Heart, Lung, and Blood Institute. The qualitative results are not changed and the dummies for the granting institute do not appear to have a significant impact. The picture for the estimated probabilities, reported in Table 4.7 conditional on the presence of journal publication and for three representative granting institutes, is also very similar to the one provided in Table 4.5.

The main insight from the exercise is a confirmation of the pattern we saw emerging in the univariate analysis: some categories are convicted significantly less than others. The result is interesting per se and gives way to speculations on what is driving it. On one hand, we can explain it assuming that the high rank scientists are acquitted more often because they receive more attention, and hence more accusations, even if they do not misbehave more than other categories. However, this would sound as a plausible explanation for differences in the probability of an allegation to turn into an inquiry. Given that our analysis is conditional to the case having reached the stage of investigation, we should assume that the allegation was not clearly unsubstantiated. Another explanation would point to the higher ability of high rank scientist to exercise influence and exploit their position to influence the outcome of the inquiry. Even in this case though, it is not clear why this ability would manifest at such a later stage and not immediately, leading allegations to be dismissed directly. A further caveat is that the result proved to be robust to all the additional controls we could use but the specification of our model is still very poor. Better and more detail dataset are surely needed to assess whether this is a robust pattern or a spurious result that will fade away once we account for some key omitted variable.

5 Conclusions

The aim of this paper has been to provide a description of pattern of interest that can be identified looking into publicly available data published by the Office for Research Integrity in its Annual Reports. The caveats and the limits to which our exercise are subjected are thoroughly analyze throughout the paper and cannot be stressed enough. However, we feel that systematic exploitation of data on cases of misconduct it is a necessary step in understanding misconduct and, as a consequence, a fundamental

prerequisite to any policy dealing with the issue. We open the way in proposing a study that without taking up the question of causality, still pushes the use of public data beyond the mere statistical tabulation and try to come up with some numbers useful, at the very least, to fuel discussion. If and when more and better data will become available, there is a chance that significantly more accurate figures can be estimated following an approach similar to the one we proposed in this research.

Bibliography

Atlas M. S. (2004), Retraction policies of high impact biomedical journals, Journal of the Med Libr Assoc 92(2), pp.242-250.

Bartlett, T., Smallwood, S., (2004), Four academic plagiarists you've never heard of: how many more are out there? The Chronicle of Higher Education 51 (17), A8.

Becker G. S. (1968), *Crime and Punishment: An Economic Approach*, Journal of Political Economy 76(2), pp.169-217.

Buzzelli D. E. (1993), *The Definition of Misconduct in Science: A View from NSF*, Science 259 (January 1993), pp.584-585 and pp.647-648.

Chubin, D., (1983), Misconduct in research: an issue of science policy and practice. Minerva 23, 175-202.

David P. A.(2004), From keeping ``Nature's secrets" to the institutionalization of ``open science", in Gosh R. A. (ed.) *CODE: Collaborative ownership and the digital economy*, Cambridge MA: MIT Press.

Enders, W., Hoover, G.A., 2004. Whose line is it? Plagiarism in economics. Journal of Economic Literature 42, 487-493.

Enders, W., Hoover, G., 2006. Plagiarism in the economics profession: a survey. Challenge 49, 92-107.

Ercegovac, Z., Richardson, J.V., 2004. Academic dishonesty, plagiarism included, in the digital age: a literature review. College & Research Libraries (July), 301-318

Hackett E. J. (1994), *A Social Control Perspective on Scientific Misconduct*, Journal of Higher Education 65(3), pp.242-263.

Harold C. S., Drummond R., (2006), Research Misconduct, Retraction, and Cleansing the Medical Literature: Lessons from the Poehlman Case, Annals of Internal Medicine 144(8).

Heitor M., Conceição P. (2007), Do we need a revisited policy agenda for research integrity?… an institutional approach, Paper to be presented at the "World Conference on Research Integrity" Calouste Gulbenkian Foundation, Lisbon, Portugal 16-18 September 2007.

Lacetera N., Zirulia L. (2007), *The Simple Economics of Scientific Misconduct*, working paper.

National Institute for Health, *A Guide to the Handling of Scientific Misconduct Allegations in the Intramural Research Program at NIH*.
Available at http://www3.od.nih.gov/oma/manualchapters/intramural/3006/. Last visit 9/15/2006.

Neale A.V., Northrup J., Dailey R., Marks E., Abrams J., (2007), Correction and use of biomedical literature affected by scientific misconduct. Sci Eng Ethics, 13(1), pp.5-24.

Office for Research Integrity, *Model Policy for Responding to Allegations of Scientific Misconduct*.
Available at http://ori.dhhs.gov/documents/model_policy_responding_allegations.pdf. Last visit 9/20/2006.

Office for Research Integrity, *Recent cases summary*.
Available at http://ori.dhhs.gov/misconduct/cases

Office for Research Integrity, *Annual reports*.
Available at http://ori.dhhs.gov/publications/annual_reports.shtml

Rhodes L. J. (2004), *ORI Closed Investigations into Misconduct Allegations Involving Research Supported by the Public Health Service: 1994-2003,* ORI research paper.

Rosamond, B., 2002. Plagiarism, academic norms and the governance of the profession. Politics 22, 167–174.

"Scientific misconduct: ORI survey is flawed",Nature, 2002 Dec 19-26 (420), pp.739-40.

Stanford University, *Scientific Misconduct: Policy on Allegations, Investigations and Reporting*.
Available at http://www.stanford.edu/dept/DoR/rph/2-5.html. Last visit 8/6/2006.

"The stars who fell to Earth", Nature, 2002 Dec 19-26 (420), pp.728-29.

Schmaus W. (1983), *Fraud and the Norms of Science*, Science, Technology & Human Values 8(4), pp.12-22.

Woessner, M.C ., 2004. Beating the house: how inadequate penalties for cheating make plagiarism an excellent gamble. Political Science & Politics 37, 313-320

Zuckerman H. (1977), Deviant Behavior and Social Control in Science, in Sagarin, E. (ed.) *Deviance and Social Change: Sage Annual Reviews of Studies in Deviance*, Sage Publications, Beverly Hills.

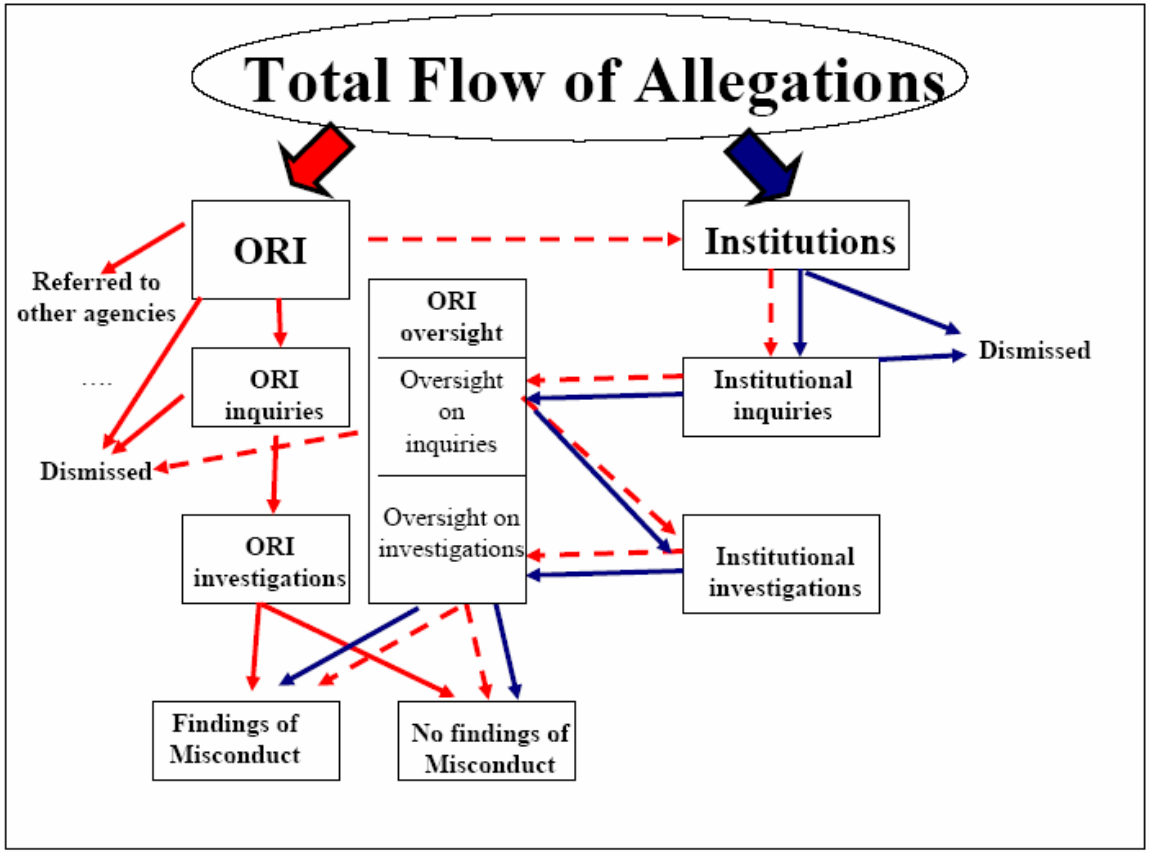


Chart 2.1: From the allegation to the final assessment of misconduct.

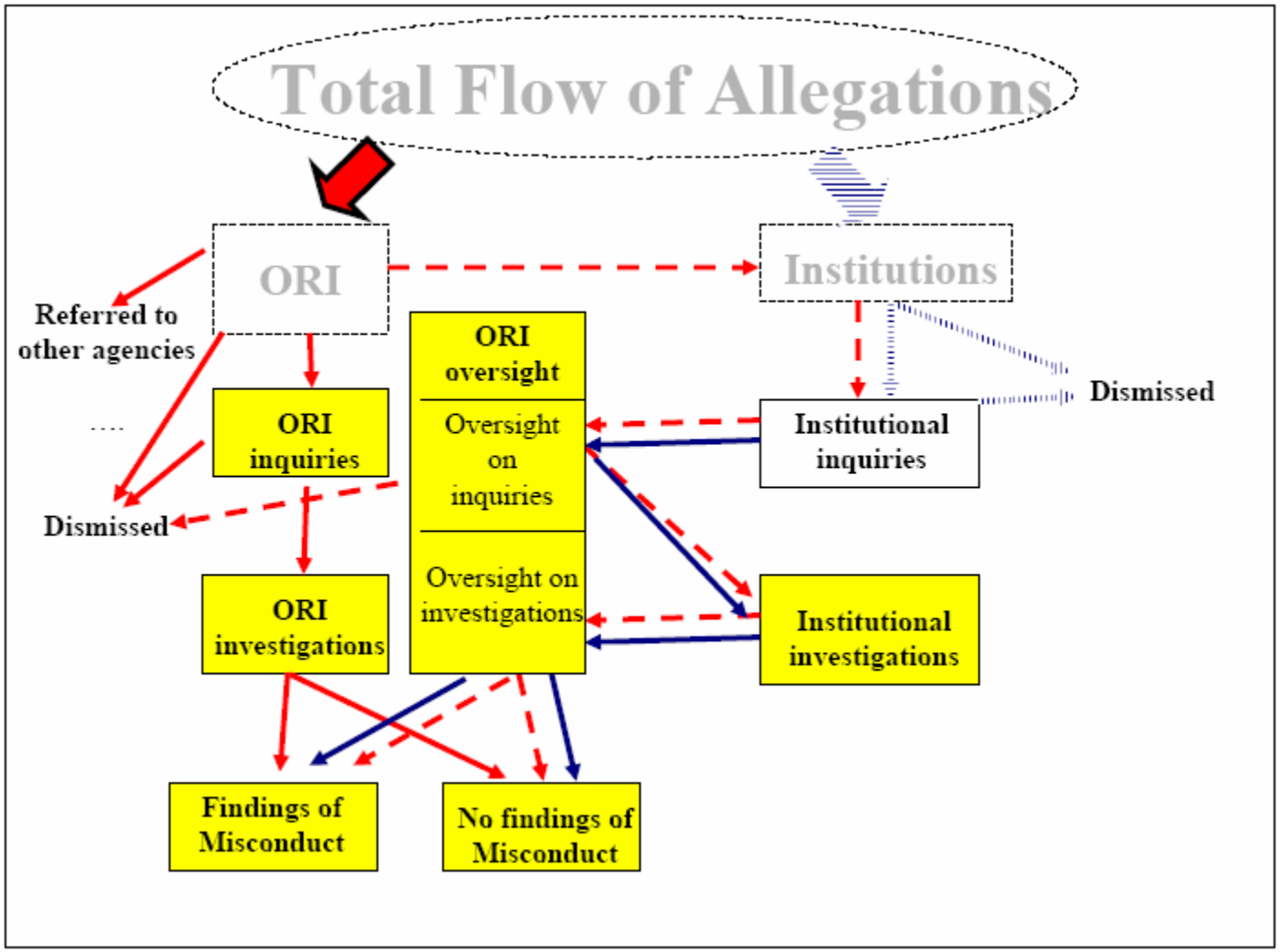


Chart 2.2: From the allegation to the final assessment of misconduct. In yellow the steps about which we can have aggregate flow information from the ORI annual reports.

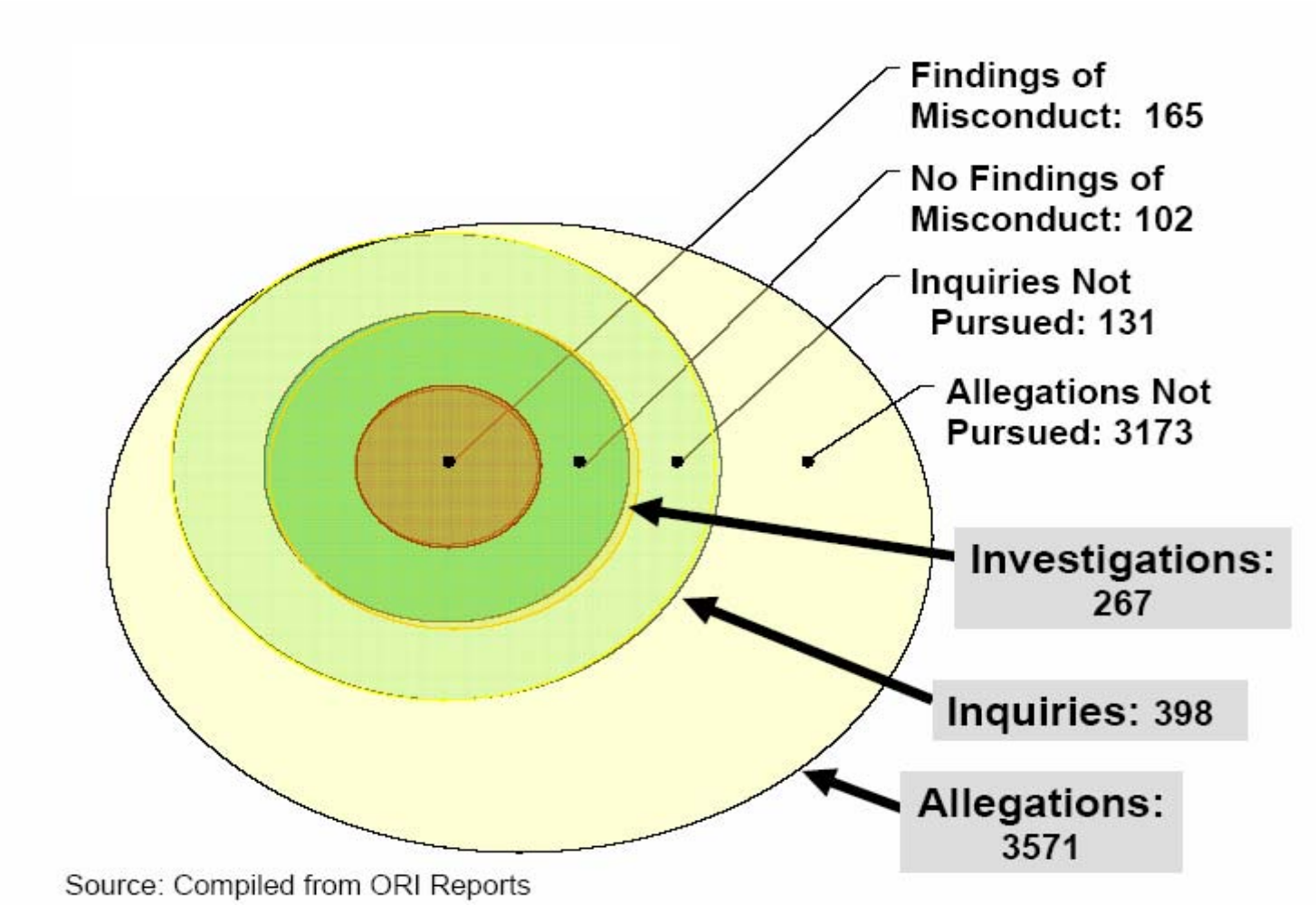


Chart 2.3: ORI case processing: totals during 1994-2005.

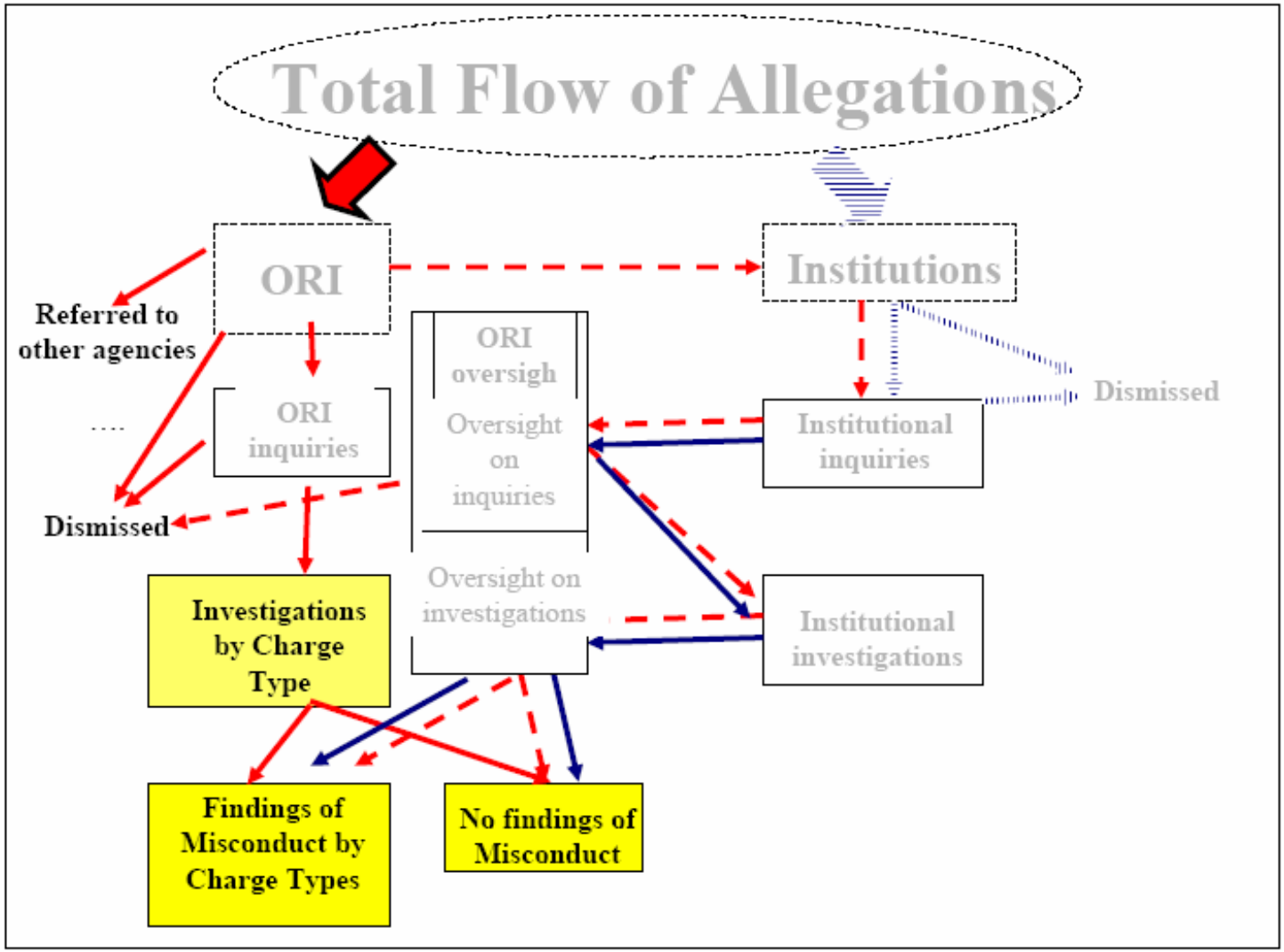


Chart 2.4: From the allegation to the final assessment of misconduct. In yellow the steps about which we can have aggregate flow information from the ORI annual reports with distinction by type of charge.

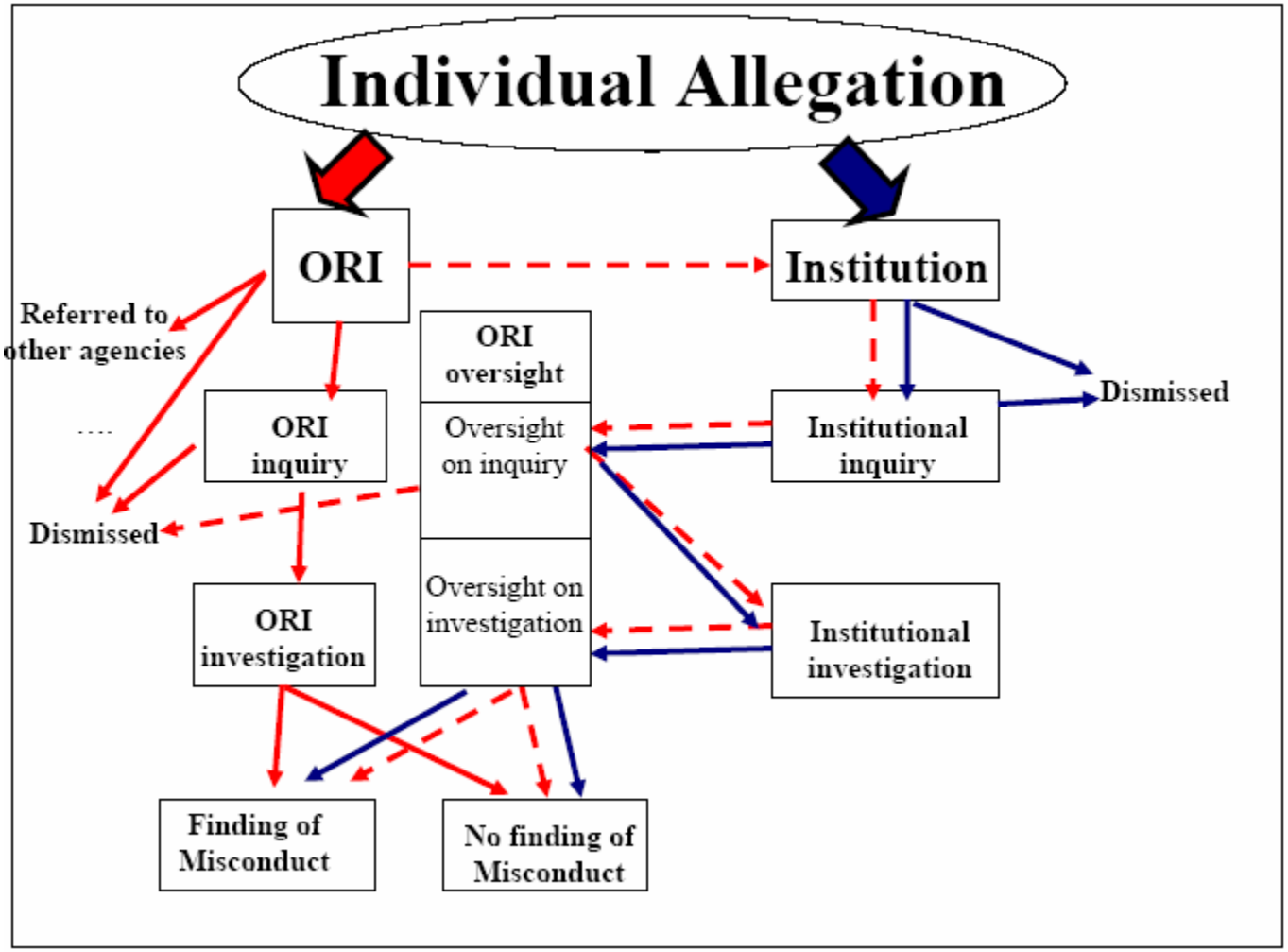


Chart 2.5: From the allegation to the final assessment of misconduct.

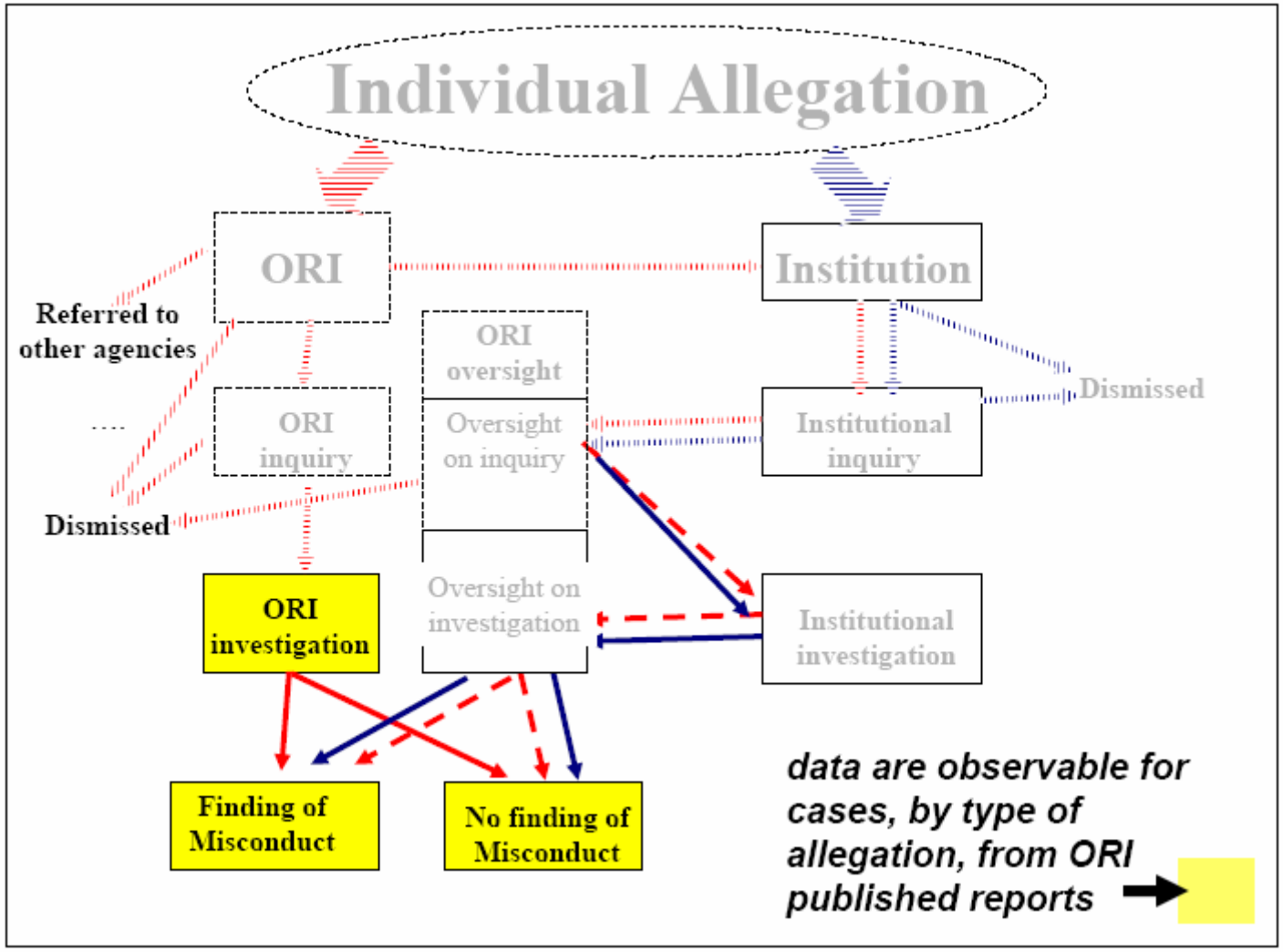
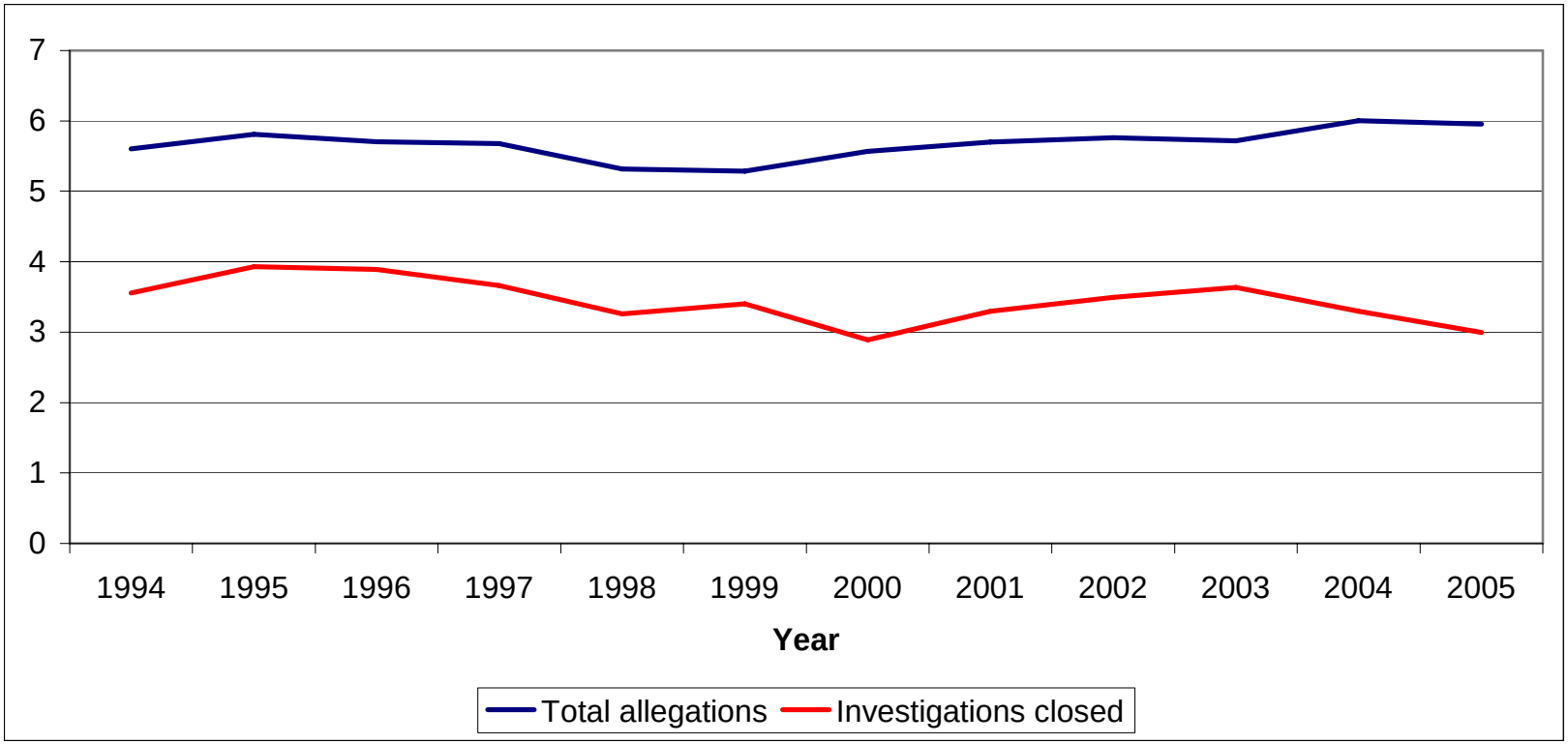


Chart 2.6: From the allegation to the final assessment of misconduct. In yellow the steps about which we have information from ORI report at the microdata level.

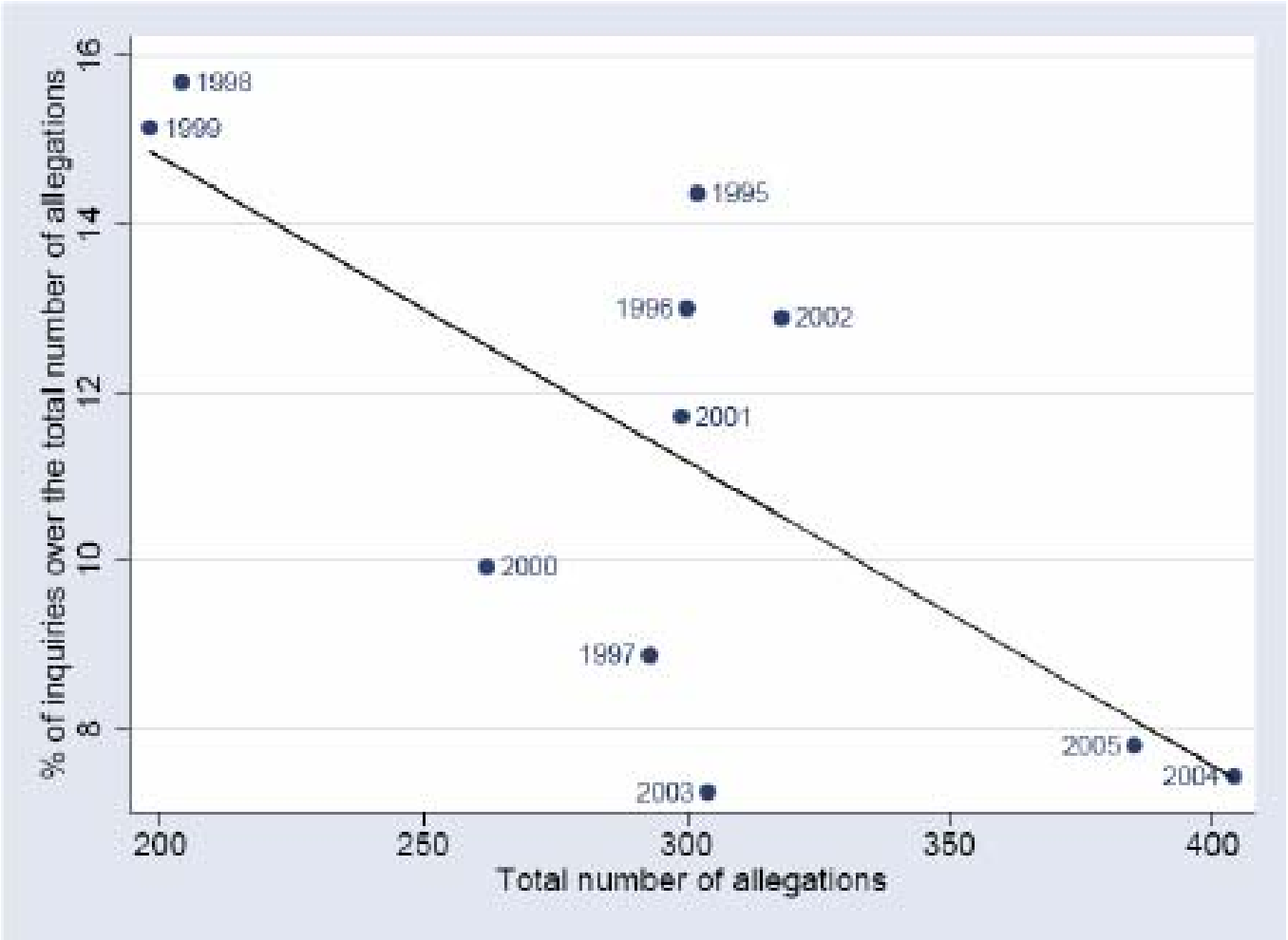


Graph 3.1: Log plot of total number of allegations and cases investigated, 1994-2005.
Source: Elaboration of ORI reports, 1994-2006.
Note: The series are plotted in logs to make them comparable.

Table 3.1: TESTS OF THE STABILITY OF MEAN ORI CASE DISPOSITION RATIOS: PRE- vs. POST 1999

ORI case disposition ratio	Mean and Standard Deviation of Annual Ratios for Period: 1994-1999	2000-2005	P-value of Difference in Mean Ratio between 1994-1999 *and* 2000-2005
Ratios to the Total Number of Allegations Received by All PHS-related Institutions of the Number of:			
Allegations reported to ORI	**0.639** (.070)	**0.622** (.083)	P > t = 0.7037 [a]
All Cases opened by ORI	**0.136** (.025)	**0.095** (.024)	P > t = 0.0161 [a]
All Investigations opened by ORI	**0.080** (.013)	**0.062** (.014)	P > t = 0.0501 [a]
Investigations closed with ORI Findings of Misconduct	**0.054** (.013)	**0.032** (.011)	P > t = 0.0098 [a]
Sequential Case Disposition Ratios:			
Allegations reported to ORI / Total Allegations	**0.639** (.070)	**0.622** (.083)	P > t = 0.7037 [a]
All cases opened / Allegations reported	**0.213** (.043)	**0.154** (.039)	P > t = 0.0322 [a]
Investigations opened / All cases opened	**0.596** (.090)	**0.664** (.138)	P > t = 0.3342 [a]
Findings of Misconduct /All Investigations	**0.704** (.202)	**0.557** (.274)	P > t = 0.3194 [a]

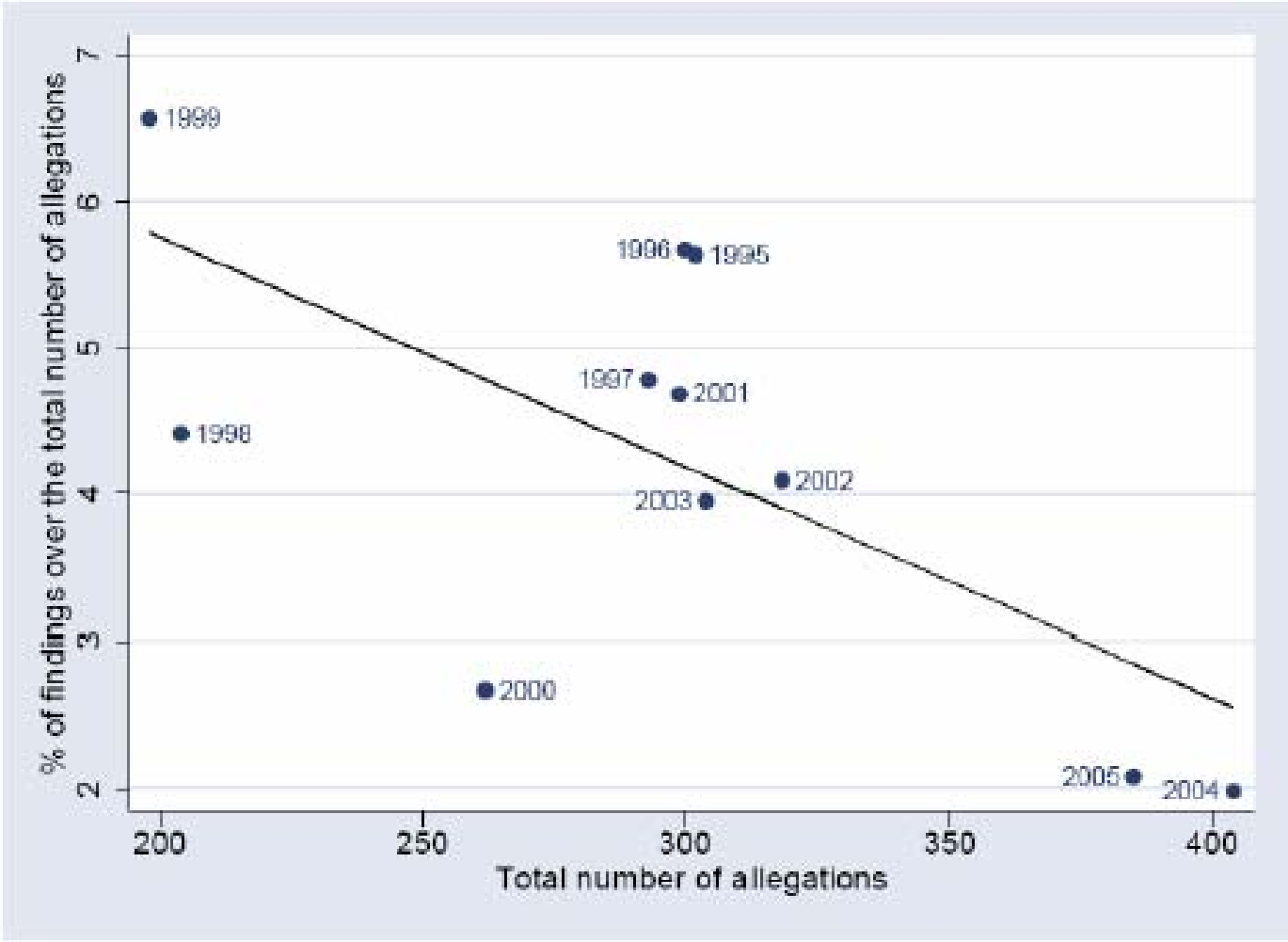
***Notes and Sources*:** See sources of annual flows underlying Pozzi-David (2007): Chart 1.3. [a] P-value of a two tails t-test of difference in means based on pooled (unequal) sample variance. Data on the total number of allegations are drawn from the "Reports on potential misconduct" that are included in the Annual Reports with one year lag. Therefore, even though we rely on Annual Reports up to 2006, data on total allegations are available only until year 2005.



Graph 3.2: Percentage of inquiries opened out of the total number of allegations plotted against the total number of allegations.
Source: Elaboration of ORI reports, 1994-2006.
Note: The black line represents the fitted values from the OLS regression whose equation line is

% of inquiries= 22.04 - .0362**Total number of allegations*, R^2=.4601.

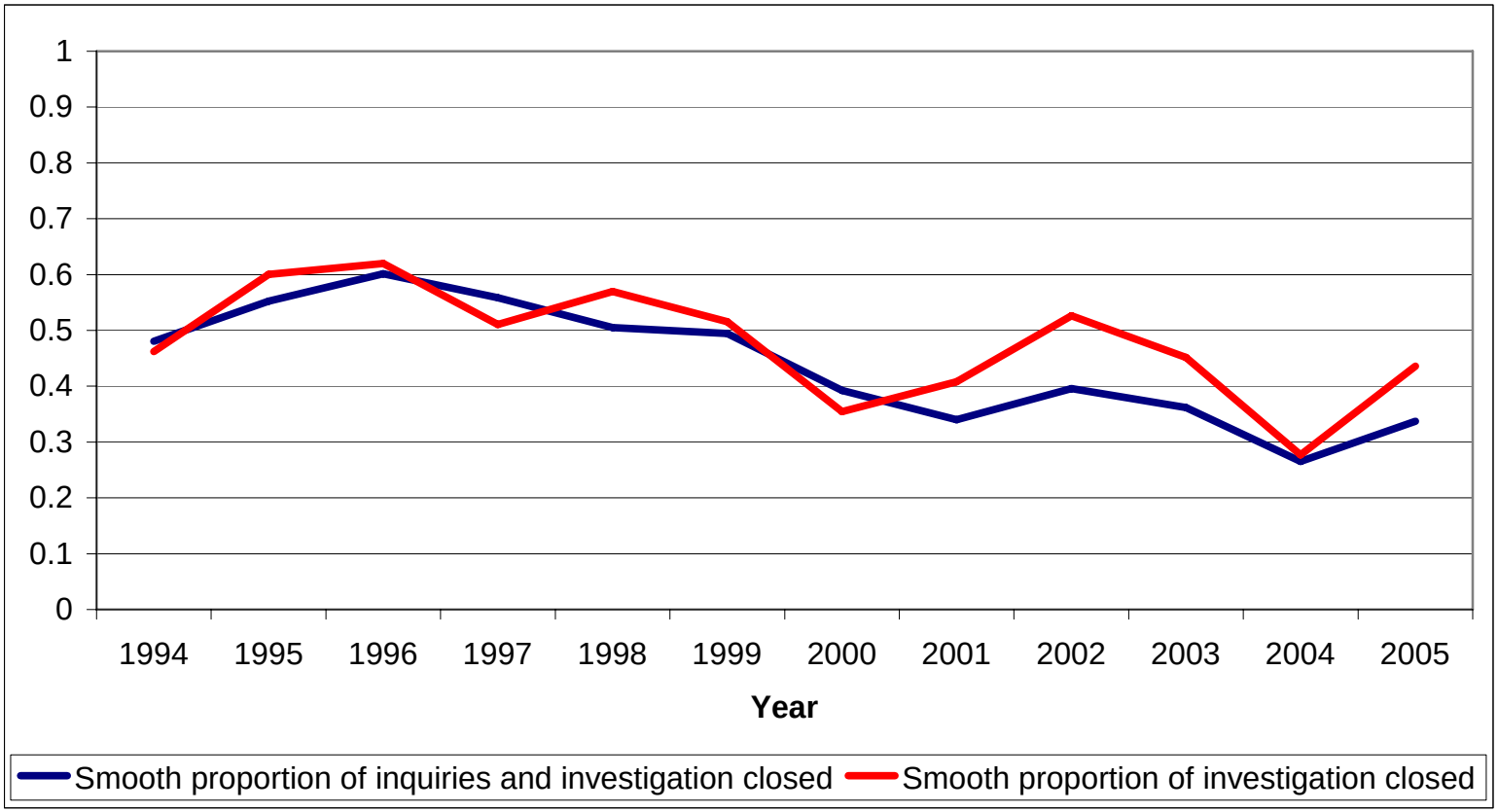
The values for year 1995 actually resulted from averaging years 1994 and 1995 to smooth any initial conditions problem.



Graph 3.3: Percentage of findings of misconduct out of the total number of allegations plotted against the total number of allegations. **Source:** Elaboration of ORI reports, 1994-2006.
Note: The black line represents the fitted values from the OLS regression whose equation line is

% of inquiries= 22.04 - .0362**Total number of allegations*, R^2=.4601.

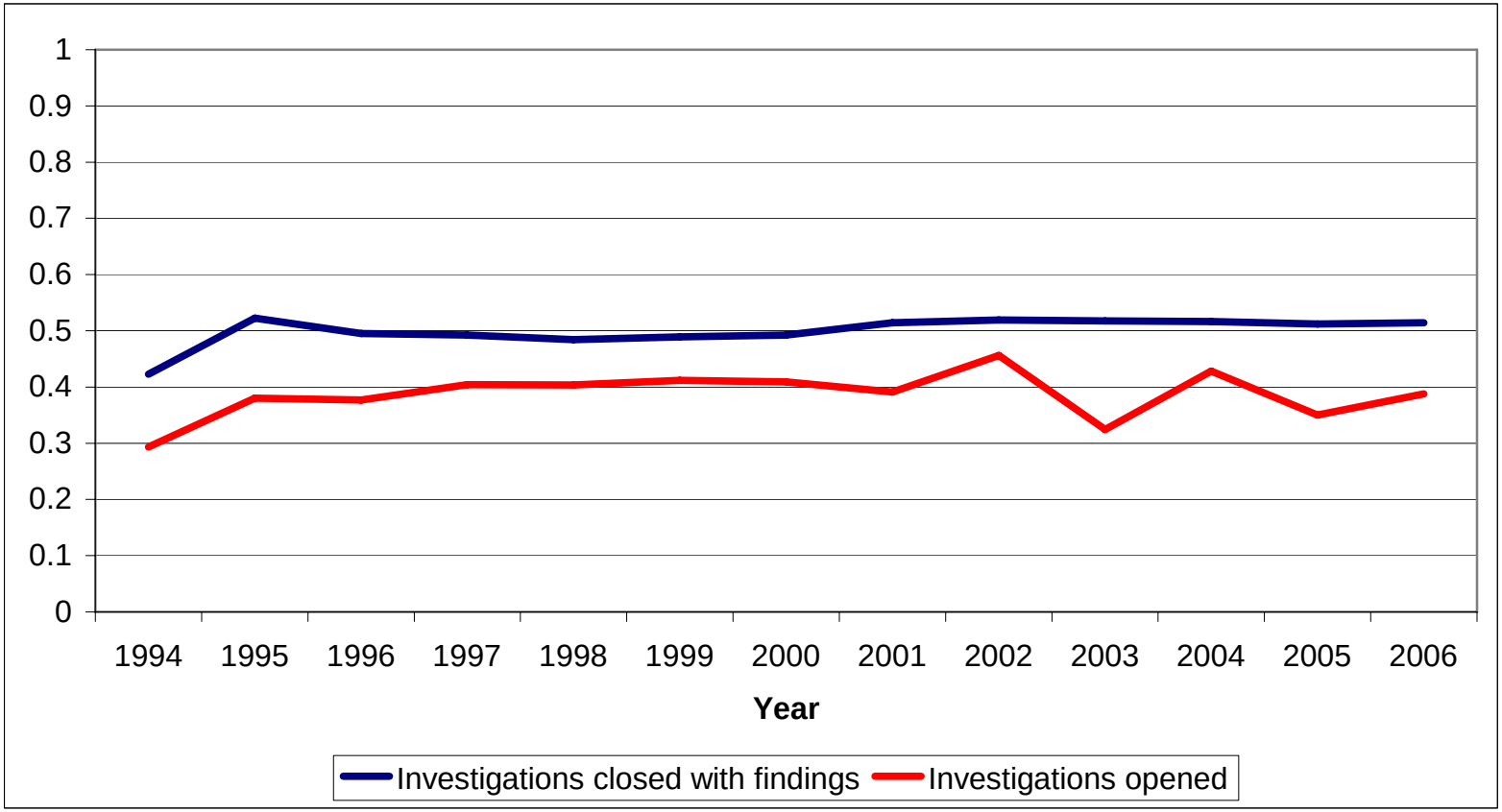
The values for year 1995 actually resulted from averaging years 1994 and 1995 to smooth any initial conditions problem.



Graph 3.4: ORI case management rate, all charges combined.

Source: Elaboration of ORI reports, 1994-2006.

Note: Smooth proportion of inquiries and investigations closed = total number of inquiries and investigations closed in a year over the sum of inquiries and investigations opened in the year and those carried to the next year. Smooth proportion of investigations closed = total number of investigations closed in a year over the sum of investigations opened in the year and those carried to the next year.
Both series are smoothed by a two years moving average.

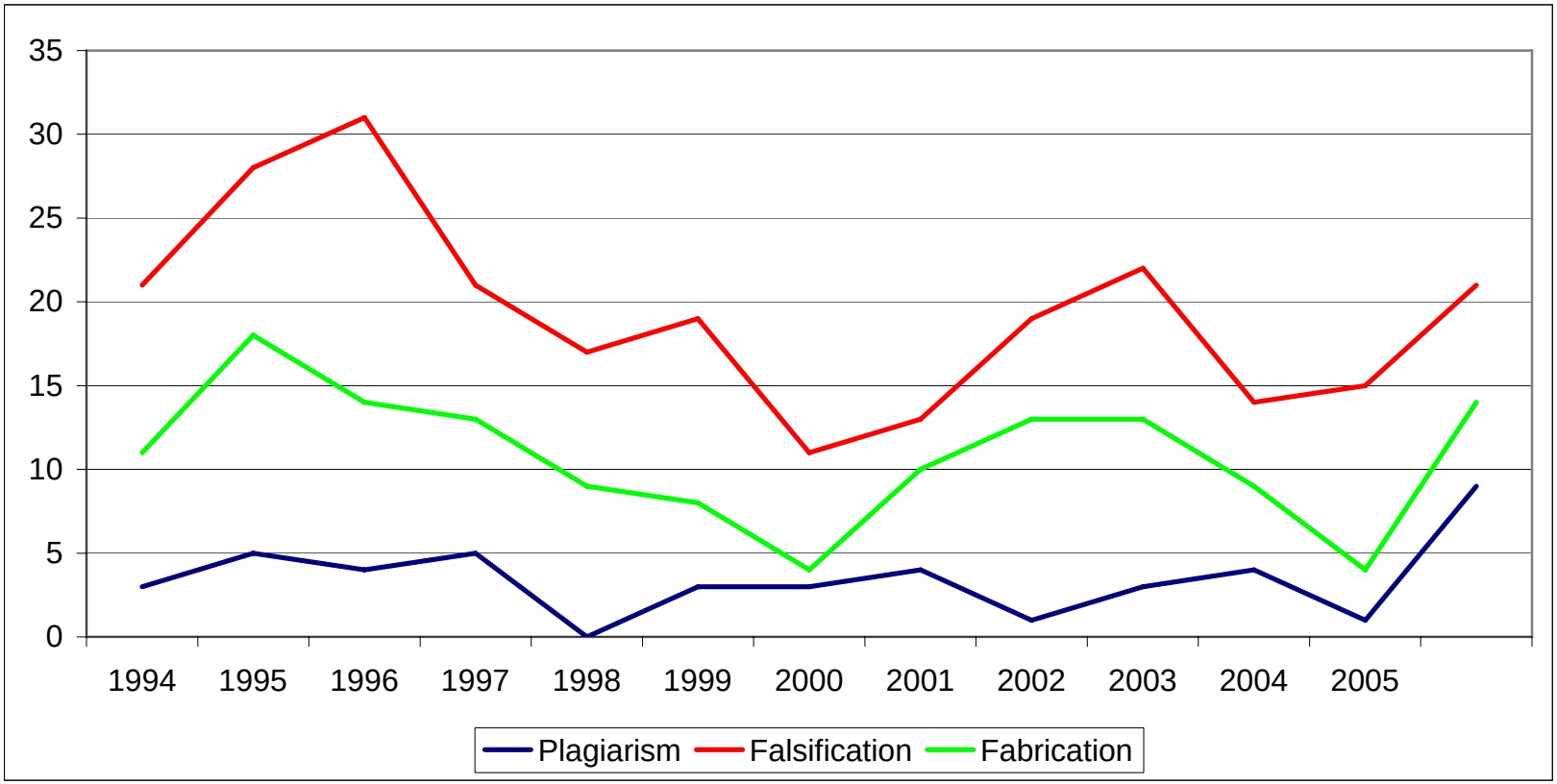


Graph 3.5: ORI case management; evolution of investigation opening and closure.

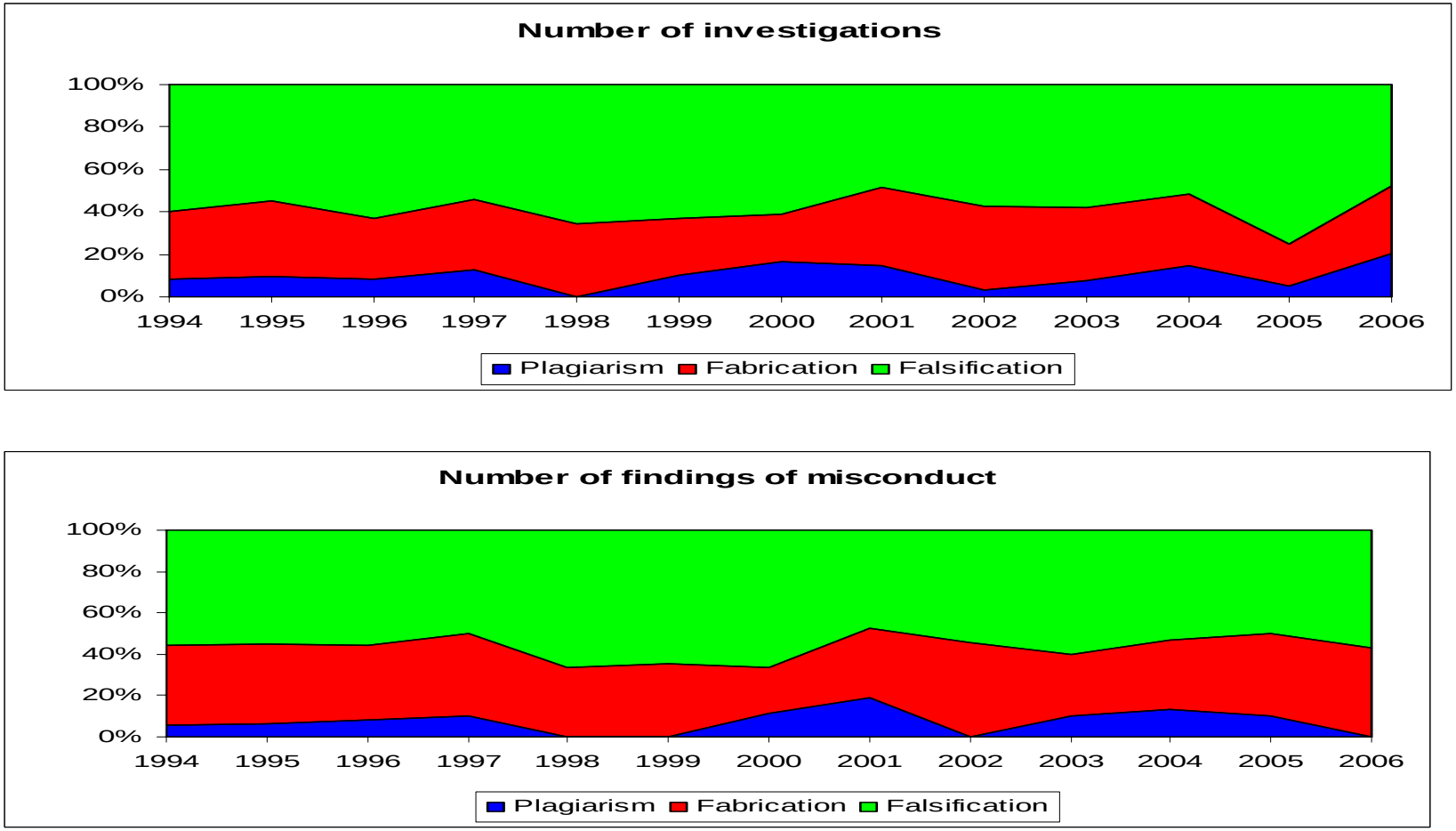
Source: Elaboration of ORI reports, 1994-2006.

Note: Investigations closed with findings= Cumulative fraction of investigations closed with findings over the total of investigations opened and carried to the next year.

Investigations opened= Cumulative fraction of new investigations opened over the total of investigations opened and carried to the next year.



Graph 3.6: Number of cases investigated by type, 1994-2006.
Source: Elaboration on ORI reports, 1994-2006
Note: Number of cases investigated = Number of investigations closed during the year.



Graph 3.7: Distribution of the number of investigations (top panel) and number of findings (bottom panel), by type of charge.
Source: Elaboration of ORI reports, 1994-2006.

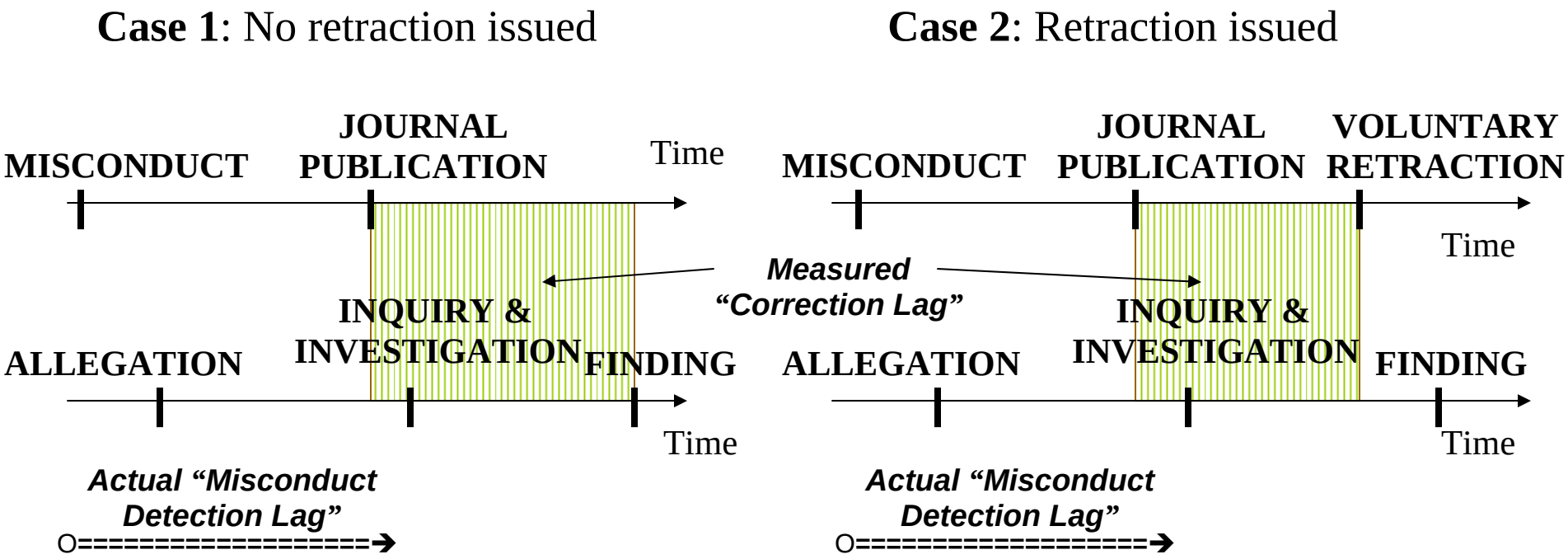
	Full professor	Associate professor	Assistant professor	Post-doc	Grad student	Research assistant	Undergrad student	Research scientist	Staff
Plagiarism	0.00	0.33	0.17	0.00	0.33	0.00	0.00	0.17	0.00
Falsification	0.21	0.18	0.12	0.07	0.08	0.02	0.00	0.15	0.17
Fabrication	0.06	0.06	0.06	0.12	0.06	0.12	0.00	0.24	0.29
Plagiarism and Falsification	0.29	0.29	0.14	0.00	0.00	0.14	0.00	0.14	0.00
Falsification and Fabrication	0.30	0.13	0.09	0.04	0.04	0.04	0.00	0.26	0.09
All three	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3.2: Distribution of charges, by position. Investigations closed without findings, 1994-2006.
Source: Elaboration of ORI annual reports.

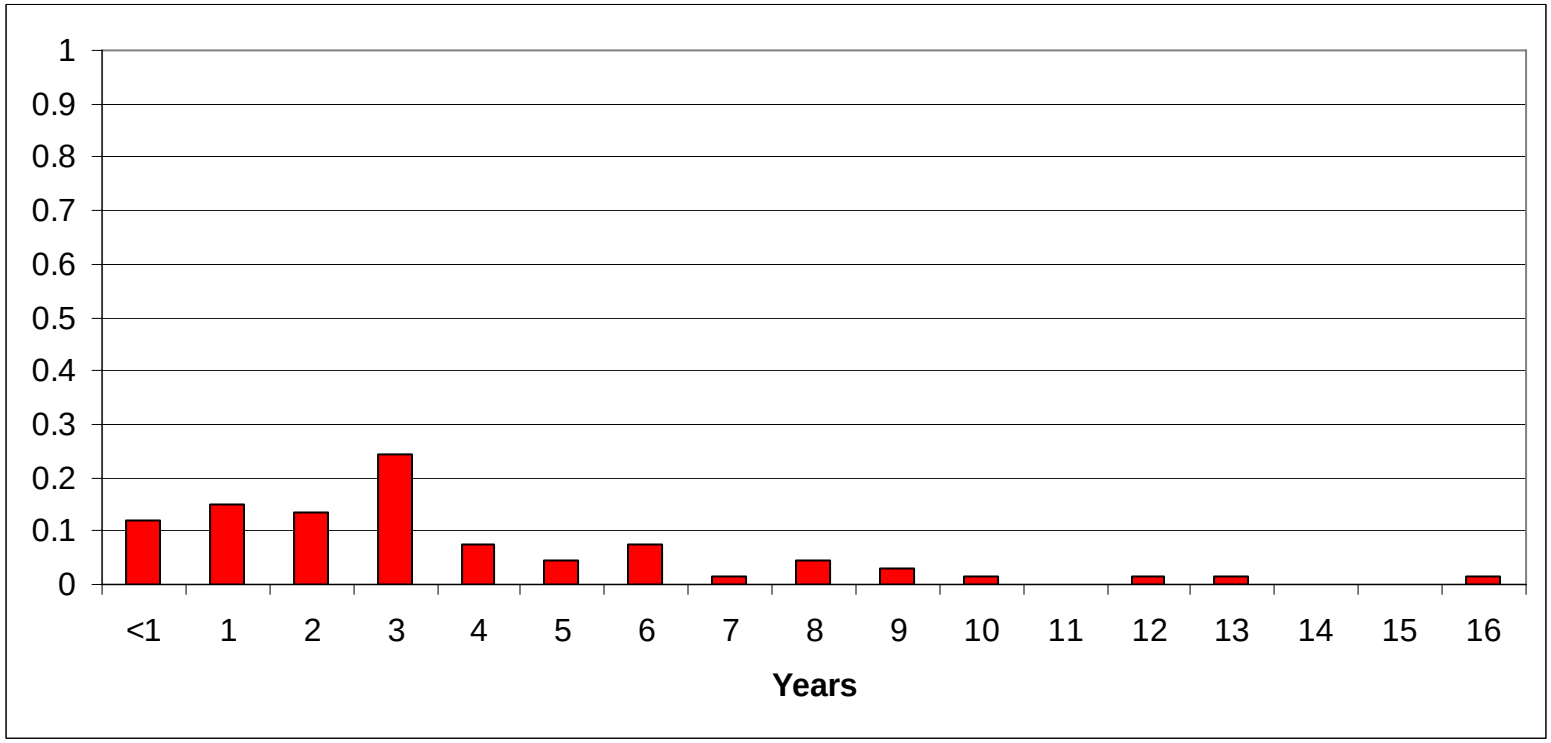
	Full professor	Associate professor	Assistant professor	Post-doc	Grad student	Research assistant	Undergrad student	Research scientist	Staff
Plagiarism	0.00	0.50	0.25	0.13	0.00	0.00	0.00	0.13	0.00
Falsification	0.08	0.08	0.10	0.19	0.19	0.00	0.00	0.15	0.21
Fabrication	0.07	0.00	0.00	0.07	0.33	0.11	0.04	0.15	0.22
Plagiarism and Falsification	0.00	0.00	0.29	0.43	0.00	0.00	0.00	0.14	0.14
Falsification and Fabrication	0.08	0.10	0.02	0.22	0.12	0.08	0.04	0.16	0.18
All three	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00

Table 3.3: Distribution of charges, by position. Investigations closed with findings, 1994-2006
Source: Elaboration of ORI annual reports

Chart 3.1: Calculation of the "Correction Lag".



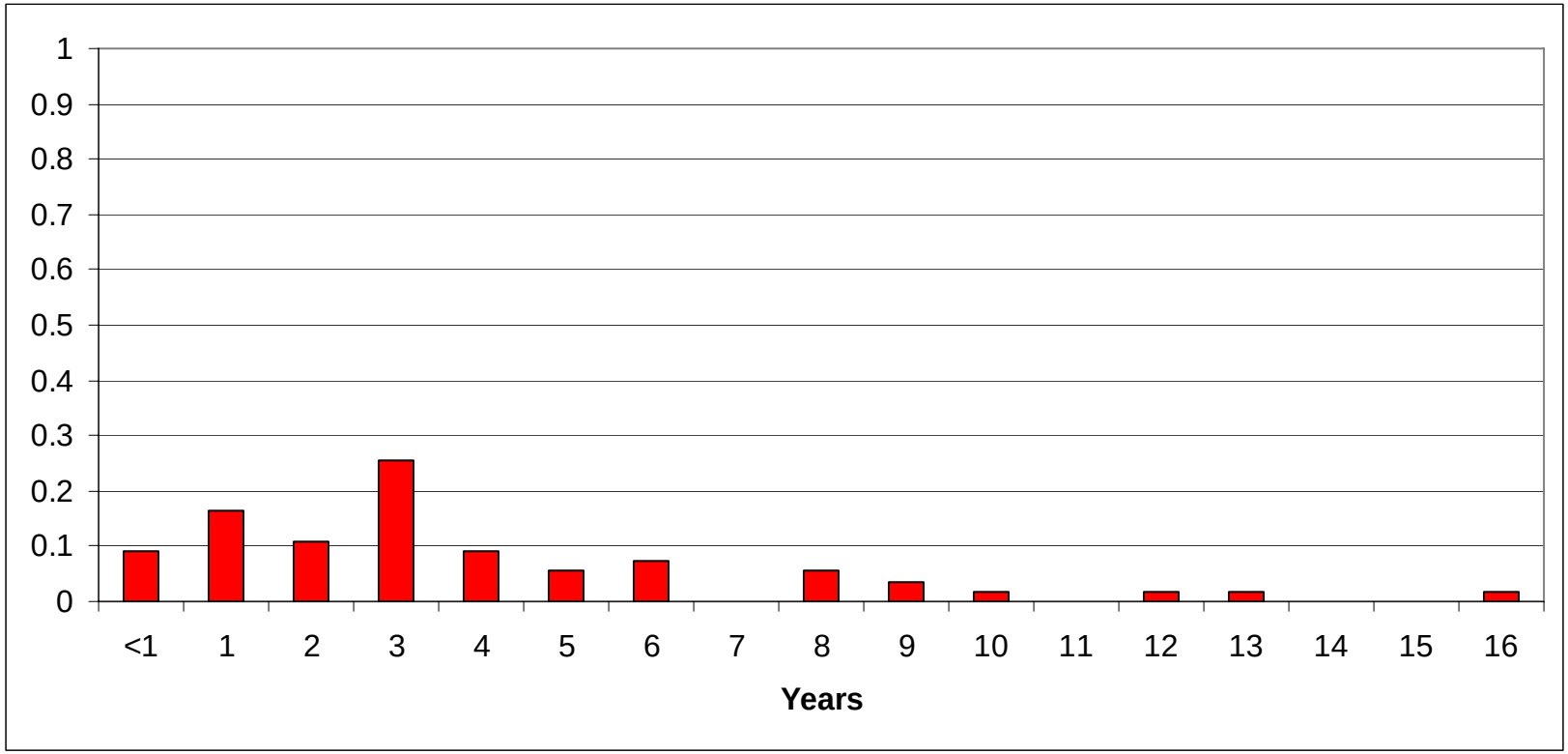
Note: The "detection lag" (between the act of misconduct and reported allegation) can be shorter or longer than the lag between journal publication and the opening of a inquiry (resulting in an investigation closed with misconduct findings). The measured "correction lag," however, overstates the lag between a post-publication allegation and the opened inquiry, as it includes the period of the inquiry and investigation, which is put at roughly 0.65 years on average (see Pozzi-David, 2007:Addendum on Inquiry Durations).



Graph 3.7: Misconduct correction lag (in years). Distribution for all charges combined, 1994-2006.

	19941994-1996	1994-1998	1994-2000	1994-2002	1994-2004	1994-2006
Mean	5.60	4.82	4.04	3.75	3.35	3.33	3.65
Median	3.00	4.00	3.00	3.00	3.00	3.00	3.00
Std. Dev.	6.11	4.35	4.01	3.68	3.24	3.06	3.32
Range	1--16	0--16	0--16	0--16	0--16	0--16	0--16
Std. Dev./Mean	1.12	0.90	0.99	0.98	0.97	0.92	0.92

Table 3.4: Evolution of the average "Misconduct correction lag" (in years), for all charges combined.
Source: Elaboration on ORI reports, 1994-2006.
Note: The detection lag is calculated only for those observations that involve publication of a paper related to the research under investigation. The date of publication of the paper is used as an estimate of the time at which misconduct took place (surely the misconduct act had already been committed prior to submission and, a fortiori, prior to the publication). As an approximation for the date of detection, we take the minimum between the date in which the investigation on the case is closed (most cases are closed within a year, so we are not too far away from the actual date of detection) and the date of eventual voluntary retraction by the respondent. In this sample, the detection date coincides with the end of the investigation in 40 cases out of 65, and we use the date of retraction statements in the other 25 cases.



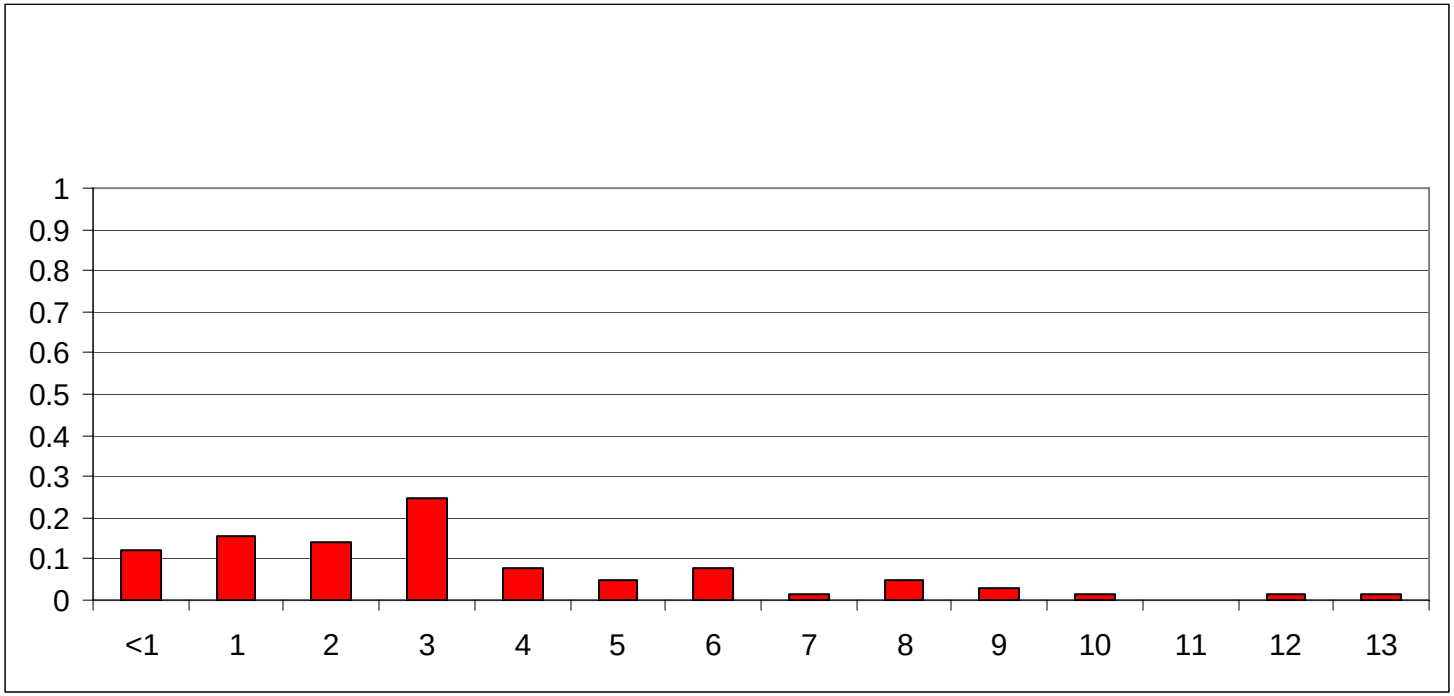
Graph 3.7: Misconduct correction lag (in years). Distribution for all cases involving falsification, 1994-2006.

	19941994-1996	1994-1998	1994-2000	1994-2002	1994-2004	1994-2006
Mean	5.60	5.00	4.26	4.10	3.52	3.54	3.91
Median	3.00	4.00	3.00	3.71	3.23	3.29	3.69
Std. Dev.	6.11	4.42	4.05	3.71	3.29	3.33	3.74
Range	1--16	0--16	0--16	0--16	0--16	0--16	0--16
Std. Dev./Mean	1.09	0.88	0.95	0.90	0.93	0.94	0.96

Table 3.5: Evolution of the average "Misconduct correction lag" (in years), for all cases involving falsification.

Source: Elaboration on ORI reports, 1994-2006.

Note: The detection lag is calculated only for those observations that involve publication of a paper related to the research under investigation. The date of publication of the paper is used as an estimate of the time at which misconduct took place (surely the misconduct act had already been committed prior to submission and, a fortiori, prior to the publication). As an approximation for the date of detection, we take the minimum between the date in which the investigation on the case is closed (most cases are closed within a year, so we are not too far away from the actual date of detection) and the date of eventual voluntary retraction by the respondent. In this sample, the detection date coincides with the end of the investigation in 40 cases out of 65, and we use the date of retraction statements in the other 25 cases.



Graph 3.8: Misconduct correction lag (in years). Trimmed distribution for all charges combined, 1994-2006.

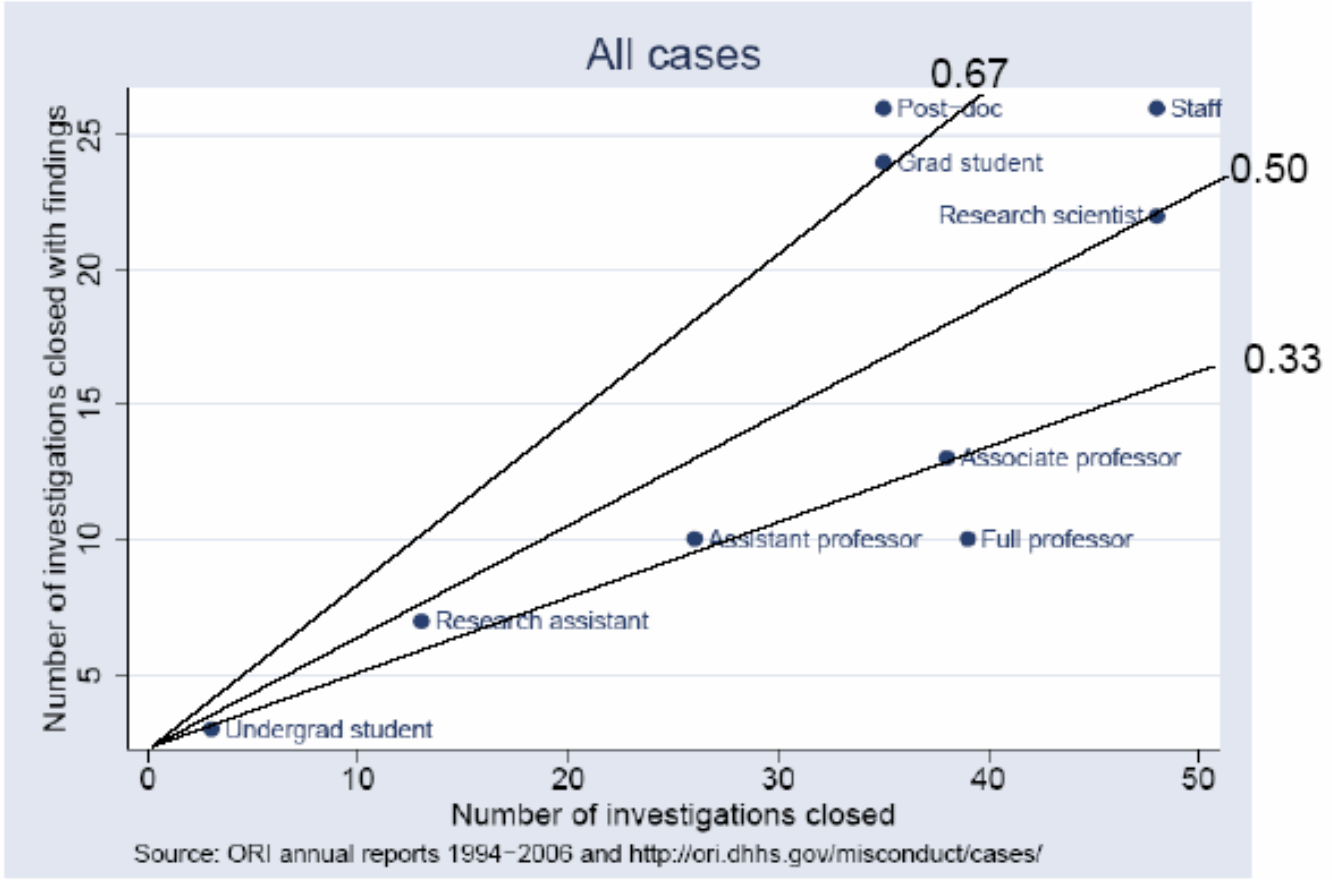
	1994	1994-1996	1994-1998	1994-2000	1994-2002	1994-2004	1994-2006
Mean	3.00	4.13	3.56	3.40	3.14	3.12	3.46
Median	2.50	3.50	3.00	3.00	3.00	3.00	3.00
Std. Dev.	2.16	3.36	3.25	3.07	2.74	2.60	2.96
Range	1--6	0--12	0--12	0--12	0--12	0--12	0--13
Std. Dev./Mean	0.72	0.82	0.91	0.90	0.87	0.84	0.86

Table 3.6: Evolution of the average "Trimmed misconduct correction lag" (in years), for all charges combined.
Source: Elaboration on ORI reports, 1994-2006.
Note: The detection lag is calculated only for those observations that involve publication of a paper related to the research under investigation. The date of publication of the paper is used as an estimate of the time at which misconduct took place (surely the misconduct act had already been committed prior to submission and, a fortiori, prior to the publication). As an approximation for the date of detection, we take the minimum between the date in which the investigation on the case is closed (most cases are closed within a year, so we are not too far away from the actual date of detection) and the date of eventual voluntary retraction by the respondent. In this sample, the detection date coincides with the end of the investigation in 40 cases out of 65, and we use the date of retraction statements in the other 25 cases.

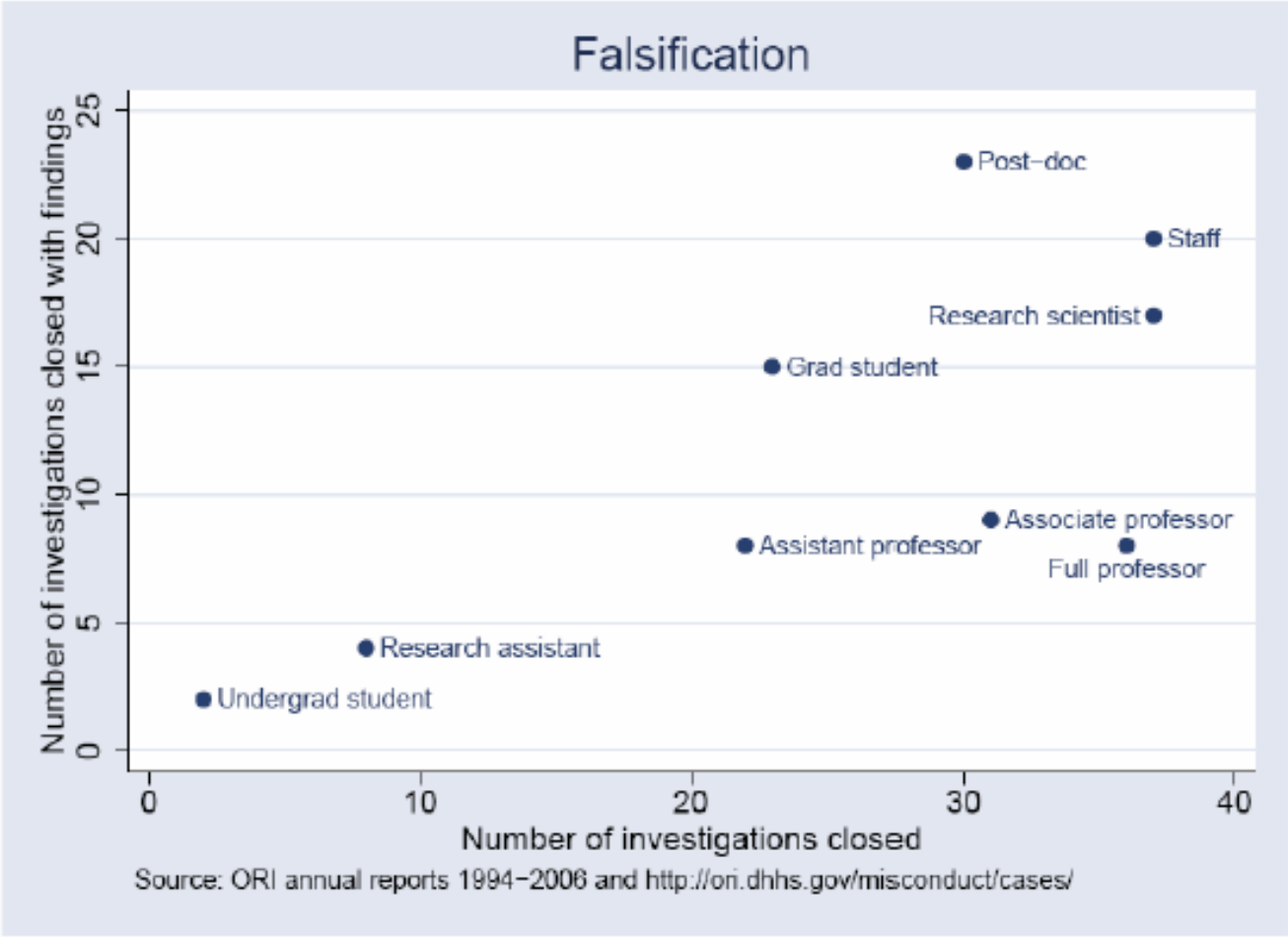
Table 4.1: Positions held by individual subjects of ORI "closed Investigations", 1994-2006.

Category	Position	Obs.	Category	Position	Obs.
Academic			*Non academic*		
Full professor		**34**	**Research scientist**		**42**
	Full professor	33		Program coordinator	6
	Faculty member	1		Project coordinator	1
				Project director	4
Associate professor		**34**		Clinic coordinator	4
				Principal investigator	3
Assistant professor		**24**		Executive manager	1
				Research associate	11
Instructor		1		Research fellow	5
				Research scientist	5
Post-doc		**32**		Scientist	2
	Post-doc	29			
	Visiting fellow	3	**Staff**		**45**
				Laboratory technician	6
Graduate student		**31**		Research technician	6
				Technician	5
Research assistant		**12**		Data manager	4
	Research assistant	9		Study counselor	2
	Assistant researcher	3		Counselor	1
				Interviewer	4
Undergraduate student		**3**		Staff	9
				Contractual employee	3
				Employee	4
				Assistant member	1

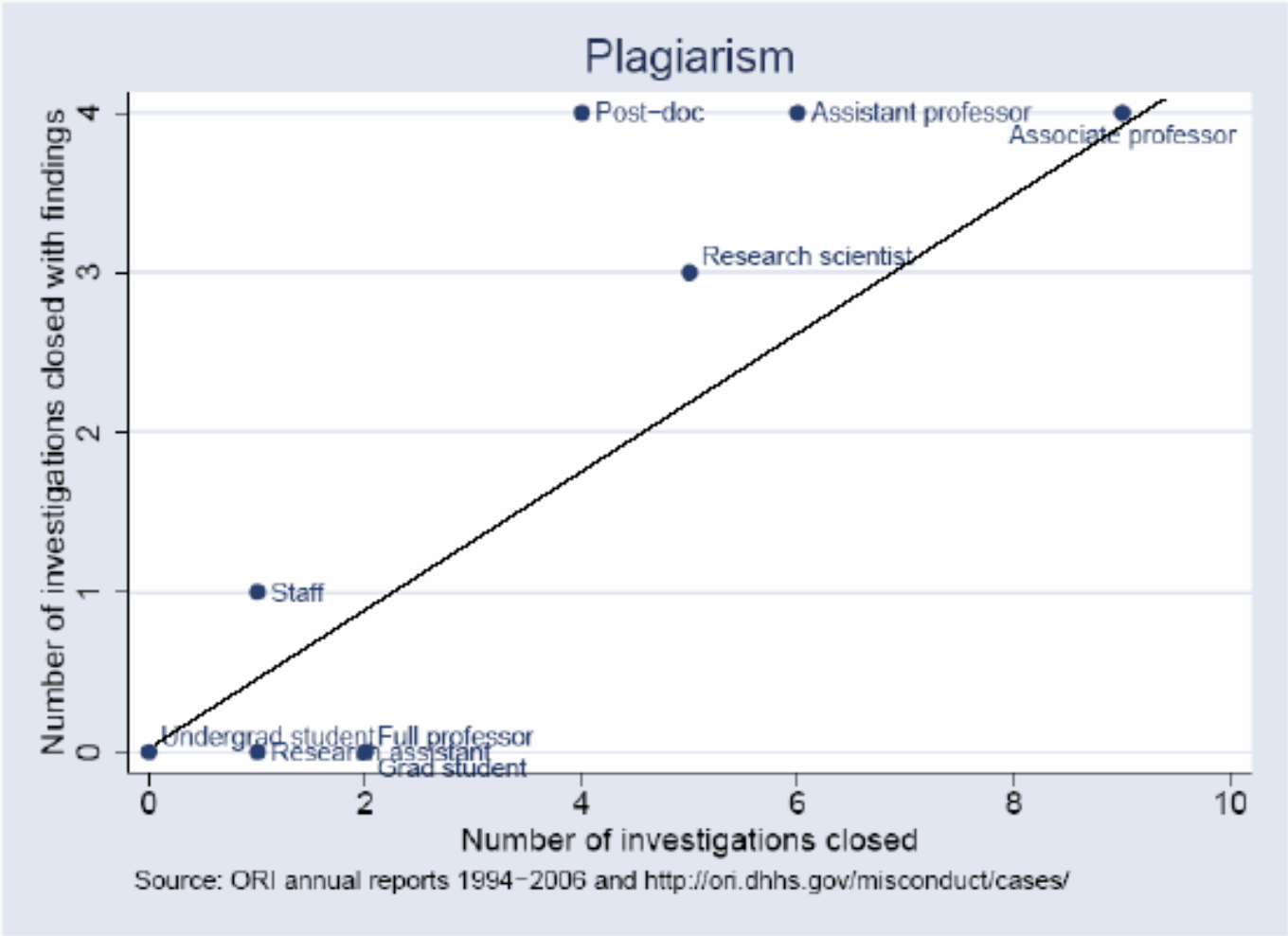
Source: ORI reports 1994-2006. The Categories in boldface are aggregates, formed for the purpose of the analysis of "position effect" on the conditional probability of finding of misconduct.
Note: The "faculty member" entry that has been absorbed into the "Full professor" category pertains to a 1997 case involving allegations of falsification and fabrication; the ORI report discussion of the case suggests that the respondent was a senior faculty member. Although the position "Instructor" is listed separately here, for the statistical analysis this case was aggregated into the category "Staff".



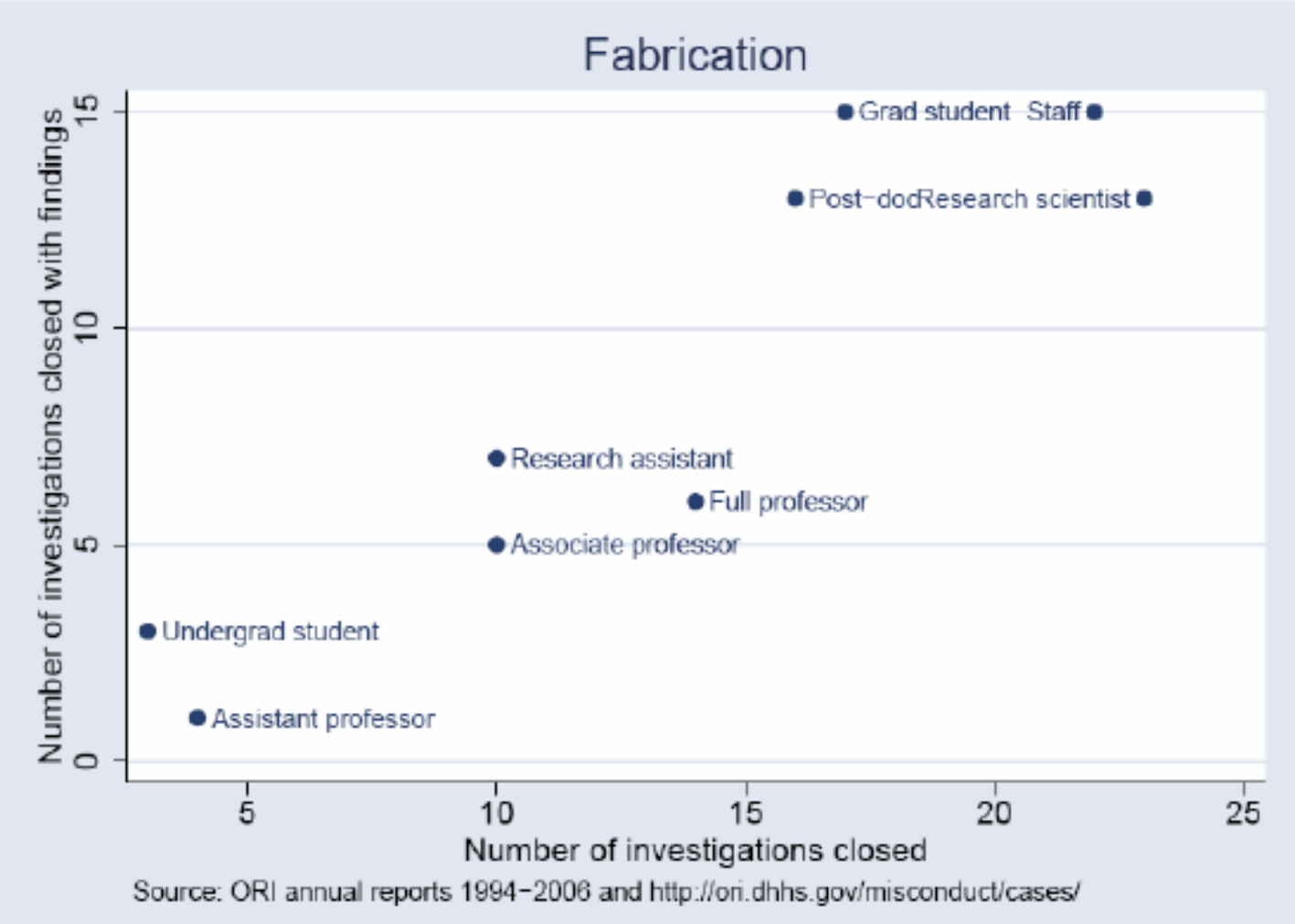
Graph 4.1: Number of investigations closed with findings of misconduct plotted against the total number of investigations closed by position held by the respondent.



Graph 4.2: Number of investigations closed with findings of misconduct plotted against the total number of investigations closed by position held by the respondent.



Graph 4.3: Number of investigations closed with findings of misconduct plotted against the total number of investigations closed by position held by the respondent.

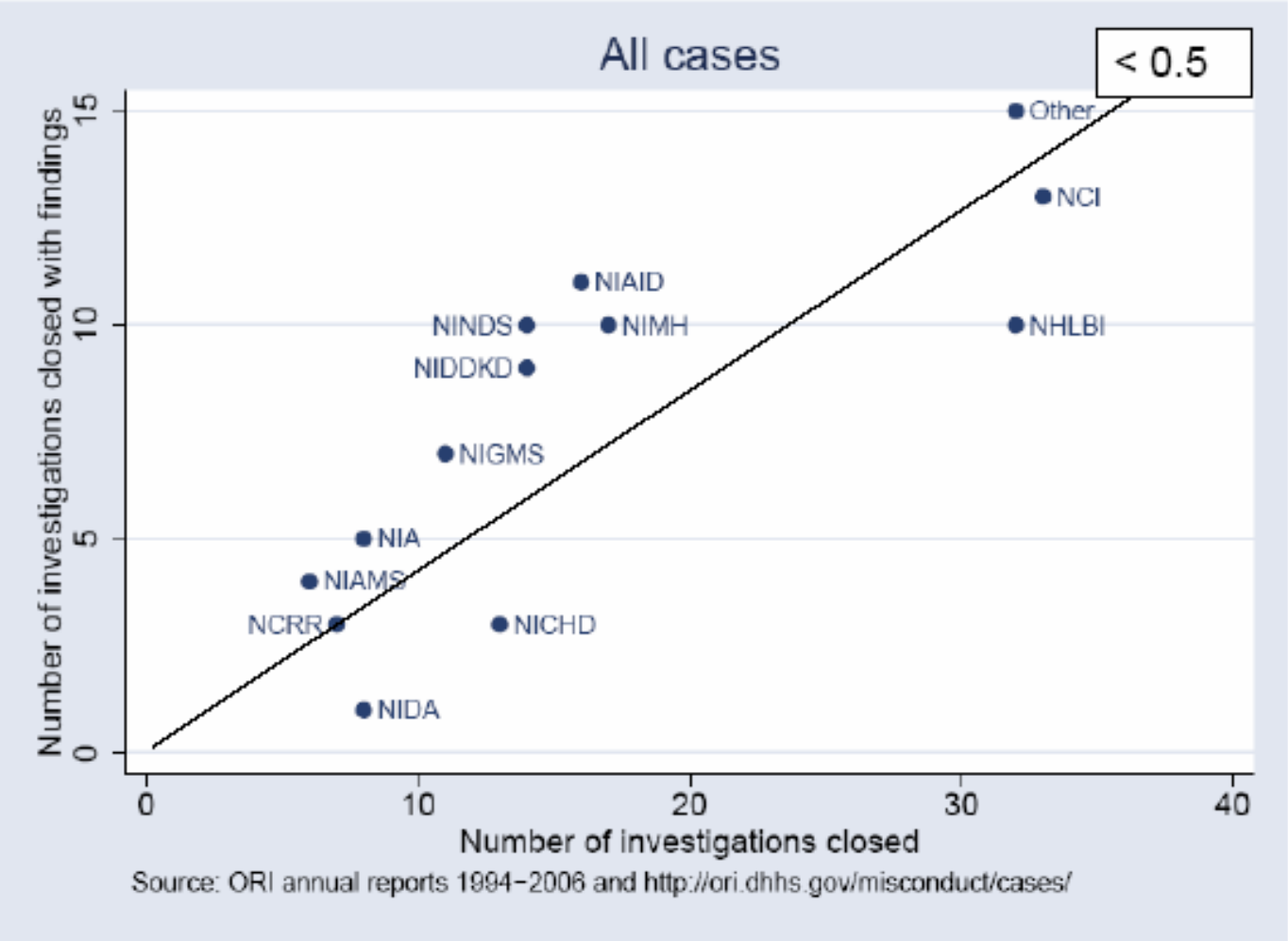


Graph 4.4: Number of investigations closed with findings of misconduct plotted against the total number of investigations closed by position held by the respondent.

Table 4.2: List of grantors, ranked by frequency of ORI "closed investigations", 1994-2006.

Institute	1994-2006 Investigations closed	1994-2006 Rank by frequency	1994-1997 Investigations closed	1998-2006 Investigations closed
Main grantors				
National Cancer Institute (NCI)	39	1.5	6	33
National Heart, Lung,and Blood Institute (NHLBI)	39	1.5	2	37
National Institute of Mental Health (NIMH)	21	3	4	17
National Institute of Allergy and Infectious Diseases (NIAID)	17	4.5	0	17
National Institute of Diabetes and Digestive and Kidney Diseases (NIKKDK)	17	4.5	2	15
Other grantors				
National Institute of Neurological Disorders and Stroke (NINDS)	15	6	2	13
National Institute of Child Health and Human Development (NICHD)	13	7	1	12
National Institute of General Medical Sciences (NIGMS)	10	8	2	8
National Institute of Aging (NIA)	9	9	1	8
National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS)	8	10.5	2	6
National Institute on Drug Abuse (NIDA)	8	10.5	1	7
National Center for Research Resources (NCRR)	7	12	0	7
National Institute of Environmental Health Sciences (NIEHS)	5	13	1	4
National Eye Institute (NEI)	4	14.5	0	4
National Institute of Dental and Craniofacial Research(NIDCR)	4	14.5	0	4
National Institute on Deafness and Other Communication Disorders (NIDCD)	3	16	1	2
National Center for Human Genome Research (NCHGR)	2	17.5	1	1
National Institute on Alcohol Abuse and Alcoholism (NIAAA)	2	17.5	0	2
Centers for Disease Control and Prevention (CDCP)	1	21.5	1	0
National Institute of Nursing Research (NINR)	1	21.5	0	1
National Institute for Occupational Safety and Health (NIOSH)	1	21.5	0	1
Substance Abuse and Mental Health Services Administration (SAMHSA)	1	21.5	0	1
Small Business Innovation Research (SBIR)	1	21.5	0	1

Source:Micro-level dataset derived from elaboration of ORI reports, 1994-2006.
Note: Investigations related to funds granted by two(three) institutes are counted twice(three times), once for each institute. Kolmogorov-Smirnov test for the equality of the cumulative distribution of closed inquiries in the two subsamples 1994-1997 and 1998-2006 gives a D=.2308, with P-value=.811. A test for the independence of the rankings in institutions by frequencies in the two sub-samples gives a Kendall Score (allowing for ties) of 40, with a P-value=.009. Both the Kolmogorov-Smirnov and the Kendall test were performed for the subset of institutes with a positive number of closed investigations in both periods.



Graph 4.5: Number of investigations closed with findings of misconduct plotted against the total number of investigations closed by grantor institute.

Table 4.3: Probit analysis of marginal effect of "Position" other than "Full professor" on the conditional probability of a misconduct finding, comparing three types of charges, 1994-2006.

	Falsification		Fabrication		Plagiarism	
Variable	**Coeff.**	**P-value**	**Coeff.**	**P-value**	**Coeff.**	**P-value**
Post-doc	.4427	.002	.1138	.720	(a)	(a)
Graduate student	.3902	.006	.4663	.049	(a)	(a)
Staff	.2878	.021	.2268	.415	(a)	(a)
Research scientist	.2197	.102	.1549	.586	.9851	(b)
Assistant professor	.1725	.237	(a)	(a)	.9961	.000
Associate professor	.0156	.908	(a)	(a)	.9945	.000
Research assistant	(a)	(a)	.2496	.381	(a)	(a)
Journal publication involved	.1937	.011	.3133	.180	.000	1
Observations	169		45		15	

Source: Probit regression analysis of micro-level dataset derived from elaboration of ORI reports, 1994-2006.

Note:*Full professor* is the omitted dummy variable for "Position".

The dummy variable for *Journal publication involved* takes the value =1 when the ORI case discussion refers to a journal article that built upon the alleged infraction and was published prior to the ORI's finding of misconduct; otherwise the "*Journal*" dummy takes value =0, which includes case of infractions in working papers, grant proposals, and journal submissions.

(a) Regressor omitted from the estimating equation to avoid perfect collinearity

(b) P-value could not be computed, due to insufficient degrees of freedom (too few observations).

Table 4.4: Predicted probabilities of a misconduct finding, conditional on the respondent's "Position" for each specified class of allegation investigated by ORI, 1994-2006.

	No journal publication involved			Journal publication involved		
	Falsification	**Fabrication**	**Plagiarism**	**Falsification**	**Fabrication**	**Plagiarism**
Post-doc	.5004	.4085	(a)	.5482	.7915	(a)
Graduate student	.4411	.8907	(a)	.6551	.9885	(a)
Staff	.3393	.5455	(a)	.5528	.8764	(a)
Research scientist	.2766	.4565	.5	.4820	.8248	.5
Assistant professor	.2374	(a)	.6666	.4335	(a)	.6667
Associate professor	.1295	(a)	.6666	.2805	(a)	.6667
Full professor	.1240	.2879	6.02E-10	.2658	.6857	6.01E-10
Research assistant	(a)	.6	(a)	(a)	.9026	(a)
Observations	169	45	15	169	45	15

Source: Probit regression analysis of micro-level dataset derived from elaboration of ORI reports, 1994-2006.
Note:*Full professor* is the omitted dummy variable for "Position".
The dummy variable for *Journal publication involved* takes the value =1 when the ORI case discussion refers to a journal article that built upon the alleged infraction and was published prior to the ORI's finding of misconduct; otherwise the "*Journal*" dummy takes value =0, which includes case of infractions in working papers, grant proposals, and journal submissions.
(a) Regressor omitted from the estimating equation to avoid perfect collinearity
Because our sample of microdata only contains investigations, all the probabilities reported have to be considered conditional on the case being investigated.

Table 4.5: Probit analysis for the marginal effect of positions other than "Full professor" and institutes other than NHLBI on the probability of finding of misconduct. **Falsification charges, 1998-2006.**

Variable	Coeff.	P-value
Post-doc	.5385	.004
Graduate student	.5336	.007
Staff	.3304	.055
Research scientist	.1285	.541
Assistant professor	.2332	.220
Associate professor	.1015	.569
Journal publication involved	.2875	.004
NIAID	.1570	.523
NIMH	.0528	.839
NIDDKD	.0109	.968
NCI	-.0886	.621
Other	-.9601	.000
Two grantors	-.2052	.289
Three grantors	.95	(b)

Source: Probit regression analysis of micro-level dataset derived from elaboration of ORI reports, 1998-2006.

Note:*Full professor* is the omitted dummy variable for "Position". *NHLBI* is the omitted dummy variable for "Institute"

The dummy variable for *Journal publication involved* takes the value =1 when the ORI case discussion refers to a journal article that built upon the alleged infraction and was published prior to the ORI's finding of misconduct; otherwise the "*Journal*" dummy takes value =0, which includes case of infractions in working papers, grant proposals, and journal submissions.

(a) Regressor omitted from the estimating equation to avoid perfect collinearity

(b) P-value could not be computed, due to insufficient degrees of freedom (too few observations).

Table 4.6: Predicted probabilities of a Finding of Misconduct in investigations of Falsification, conditional on respondent's Position and selected grantor Institute, 1998-2006.

Institute	Position	No journal publication involved	Journal publication involved
	Post-doc	.3487	.6888
	Graduate student	.3432	.6834
	Staff	.3393	.5455
NHLBI	Research scientist	.1724	.4746
	Assistant professor	.1121	.3690
	Associate professor	.0587	.2467
	Full professor	.0308	.1617
	Post-doc	.5234	.8264
	Graduate student	.5174	.8224
	Staff	.3095	.6494
NIAID	Research scientist	.1477	.4343
	Assistant professor	.2212	.5450
	Associate professor	.1317	.4062
	Full professor	.0776	.2946
	Post-doc	.2443	.5748
	Graduate student	.2396	.5688
	Staff	.1059	.3566
NCI	Research scientist	.0361	.1796
	Assistant professor	.0643	.2616
	Associate professor	.0308	.1614
	Full professor	.0149	.0983

Source: Elaboration on ORI reports, 1998-2006. **Note:**The three Institutes for which predicted probabilities are displayed were selected as representative of the range of variation in effects among the major grantors. Corresponding probability estimates for all the institutes specified in the marginal effects model [your Table 5] are available on request. Because our micro-dataset only contains observations on the outcomes of closed investigations, the probabilities reported here are conditions on the cases having been selected for investigation. The dummy variable for *Journal publication involved* takes the value = 1 when the ORI case discussion refers to a journal article that built upon the alleged infraction and was published prior to the ORI's finding of misconduct; otherwise the "*Journal*" dummy takes the value = 0, which includes case of infractions in working papers, grant proposals, and journal submissions.

Appendix A

A Comparison and Reconciliation of the Pozzi-David Dataset with the Statistics Reported by Rhoades`(2004) Study of ORI Investigations Closed during 1994-2003

A number of discrepancies appear between the annual totals and cumulative figures for the ORI case flow recorded in the dataset that was prepared for study, and the statistics presented by the a comprehensive report made by a member of the ORI research staff, Lawrence J. Rhoades' and released in 2004: "ORI Closed Investigations into Misconduct Allegations Involving Research Supported by the Public Health Service: 1994-2003" [abstracted in the ORI Annual Report for 2004: p. 27, and available at http://ori.hhs.gov/research/intra/documents/Investigations1994-2003-2.pdf.

Divergences and Reconciliations

In the following we report the results of proceeding "table-by-table" through Rhoades' publication and undertaking to reconcile or otherwise account for such discrepancies that have been found by comparing the numbers shown there and the corresponding figures from the dataset from which we have worked. We number and highlight (in boldface) the successive tables in Rhoades (2004) that have been reviewed for this purpose:

1. Rhoades' discussion accompanying **Table 1** remarks that

> "From 1994 to 2003, research institutions and the Office of Research Integrity (ORI) conducted 259 formal investigations into allegations."

We agree perfectly on this statement [see Pozzi-David Dataset (2007), Table 1 of file: time_series_analysis_table_v2, column "cases closed/investigations"].

2. But Rhoades' text then continues (pp. 2-3) as follows:

> "Over the ten-year period, ORI received 1,777 allegations. Nineteen percent (329) of these allegations resulted in new ORI misconduct cases because they met the three conditions required to establish PHS jurisdiction: (1) the alleged behavior fit the definition of research misconduct in the 1989 regulation; (2) the research involved was supported by the PHS; and (3) the allegation contained sufficient information to permit the allegation to be pursued. Twelve percent (218) were referred to other federal agencies that had jurisdiction over the alleged misconduct. No action was possible by ORI on the remaining allegations (69 percent) because they did not meet the conditions required to establish PHS jurisdiction. ORI also took no action on some allegations because they were handled by the National Institutes of Health (NIH). Annually, ORI received [4]
> an average of 178 allegations; the median was 184 and the range 112-244.

There we disagree, as we count only 1733 allegations, not 1,777, which is a not inconsiderable difference**.** What accounts for the discrepancy? Taking a closer look, it is found that for 2003 the data we find 141 allegations, whereas Rhoades (2004) reports 183 (a substantial 42 absolute and proportional discrepancy). Consulting the relevant ORI it can be seen that Rhoades' review has faithfully reported the published total for the number of allegations reported to ORI in 2003, which was 179.

When one looks more closely at the table from the ORI 2003 Annual Report (p. 3), reproduced below, however, it appears that main source of the problem resides therein:

Disposition of Allegations in ORI, 2003

Handling of allegations - outcome in ORI		*Number of allegations*
Pre-Inquiry Assessment by ORI of allegations:		38
that were made to ORI directly	27	
that were made to NIH initially	11	
No Action Possible Now or No Action		83
Referred to other Federal agencies		15
Handled by NIH (for other allegations made to NIH)		5
TOTAL		179

Something is awry with the tallying of the two sub-categories of the pre-inquiry assessments -- those made to ORI directly and made to NIH initially (and then reported): 27 and 11 plainly represent the decomposition of the 38 total shown at the top of the right-column of this table, and not *additional* allegations. Therefore, the total number of allegations for 2003 should be 38+83+15+5=141, rather than the printed figure of 179 (= 38+27+11+83+15+5). The published Report appears to have double-counted the 38 pre-inquiry assessments, and whereas Rhoades (2004) accepted the result, our dataset does not. Further confirmation of the validity of the latter (if needed) is provided by the fact that none of the ORI Annual Reports for other years repeats the 2003 Report's inclusion in the overall total of the breakdown sub-totals for the Pre-Inquiry Assessment of allegations.

3. But, the foregoing that does not account for the entire discrepancy: had we followed Rhoades (2004) **Table 1** in accepting the *published* total for Allegations in 2003, we still would have counted 1,771 allegations, which falls 6 allegations short of the grand total that his review reports. The source of the remaining discrepancy, on a year-by-year basis, would be those summarized in the table below:

Table A1: Discrepancies between Rhoades (2004) and Pozzi-David (2007) in the count of allegations reported to ORI.

Year	Pozzi-David (2007)	Rhodes (2004)	Difference
1999	129	130	1
2000	173	172	-1
2001	196	199	3
2002	191	190	-1
2003	179	183	4

Note: The Pozzi-David count for 2003 is inflated by 38 allegations to make it comparable with Rhodes (2004) due to the double-counting issue.

The difference of 6 allegations arises from several discrepancies for the years from 1999 onwards, where Rhoades (2004) sometimes exceeds and sometimes falls below the corresponding annual totals found from the Pozzi-David (2007) dataset. In every instance, however, the numbers in our dataset (save for 2003) correspond to those that were published in the published ORI Annual Report.[16]

We believe these minor discrepancies appear may reflect post-publication revisions of the figures in the ORI Reports, to which Rhoades (2004) had access. This conjecture is reinforced by the observations that some among the Annual Reports contain summary statements that recapitulate totals for previous years (i.e., while reporting the number of allegations for 2005, the number of allegations in 2004, 2003…are cited, or tabulated for purposes of comparison, and one occasion the figure mentioned for an earlier year (e.g. 2003, cited in 2005) differs from that which was published in the Annual Report for that date (e.g. 2003). Although no notes or comments accompany the variant figures on those occasions, declaring them to be official revisions, our normal practice was to accept these later, ostensibly "revised" figures as correcting the earlier report. Yet, the possibility remains that these retrospective statements in the Annual Reports are themselves erroneous.[17]

4. In Table 1 of Rhoades' (2004) the series for "No action possible" bears little correspondence with the underlying numbers compiled in our dataset. But, here again, our dataset shows 73 allegations reported for 1999 in regard to which "no action was possible," which corresponds exactly with the number published in Table 1 page 8 of the ORI Annual Report for 1999; whereas Rhoades gives 89 as the number of cases. For all the other years, his totals are above ours. In this case it is likely that Rhoades (2004) has based his series either on a variant set of figures, or is referring to an entirely different definition of cases in which there was "no action possible;" whereas we have followed the Annual Reports. Whether the

[16] Specifically: Table 1 page 8 for 1999, Table 1 page 2 for 2000, Table 1 page 3 for 2001, Table 1 page 9 or 2002, Table 1 page 3 for 2003.

[17] If it is indeed ORI's practice to correct published data when errors are discovered, it would be useful to dispel the ambiguity by periodically publishing revised retrospective time series – as is done by other government agencies, such as the U.S. Commerce Department Bureau of Economic Analysis, and Department of Labor Statistics.

suspected "alternative" series is the better of the two cannot be determined from the sources available to us, but, again, there are no comments in Rhoades (2004) accompanying text or table notes that indicate that such an alternative series was substituted for the published statistics.

5. Table 3 and 4 of Rhoades (2004) deal with a distinction between institutional inquiries/investigations and those carried on by ORI. We did not make such a distinction since ORI always carried on very few cases on its own and stopped doing that at all from 1999 onwards.

Table 5 is about the distribution of the outcomes of closed investigations, between cases with "findings" of misconduct and those where there was "no finding". Our data on this question agree almost perfectly with those presented by Rhoades. He reports 133 cases were closed with findings in 1994-2003, and we count 134. Looking year by year, two things emerge: we have one case more than Rhoades in 1999 and one less in 2000. In this case an outcome in one of those two cases may have been other than the one reported by the Annual Report for 2000, or may has been altered subsequent to it (but our dataset is at least consistent with the latter). The difference in the 1994-2003 therefore origins elsewhere, and arise from the additional "finding" that our dataset shows for 1996: 17 findings, against Rhoades' report of 16. But Table 6 page 28 of the ORI Annual Report for 1996 gave 17 investigations closed with findings of misconduct.

6. Tables 7 and 8 in Rhoades provide a breakdown by type of misconduct. Here the discrepancies are more readily accounted for, since it is clear that Rhoades is reporting case numbers, rather than numbers of allegations. The former practice maintains the distinctions among single and multiple allegation cases, reporting plagiarism, falsification, fabrication, and "falsification/fabrication", "falsification/plagiarism", and so on, as distinct types. We have counted the number of individual allegations by type, which double-counts the number of "cases" (and hence respondents) when there are multiple allegations. Thus, we count only the allegations of falsification, fabrication and plagiarism and, whenever there is a case that involves falsification/fabrication our procedure adds one unit each to the totals for fabrication and for falsification.

7. Tables 9-12 in Rhoades (2004) deal with distinction between intra-mural and extra-mural research**; Tables 13-16** with institutional settings, and **Tables 17-18** with the funding mechanism. Since we did not tabulate any of this data, no reconciliation check can be made for these data.

8. Table 19-22 in Rhoades gives counts of respondents in misconduct investigations by academic position. It could be compared to the data shown in Table 3.2. There are small discrepancies even in this case but, again, these are readily accounted for. Our dataset did not compile totals by respondents' academic/ non-academic rank from ORI aggregate figures (as was the case for the annual flows and cumulative total in the preceding tables). Rather the "position" information was individually extracted and entered by hand from the texts of the published reports on closed investigations. This allowed the "positions data" to be used in the micro-level analysis that estimated conditional mean probabilities of misconduct findings.

Discrepancies between the totals based on the micro-data and Rhoades' figures Table 19-22 therefore are traceable to a number of procedural differences: (i) Cases that were not relevant to the charges of misconduct (plagiarism, falsification and fabrication) were not tabulated as they were not relevant for the intended analysis (e.g. people mis-using funds or harassing lab mates); although these were infrequent, it would appear that they would be included in the totals given by Rhoades. (ii) From the published summaries of close investigation was not always possible to discern the respondent's precise academic position (especially in the reports during the mid-1990s) in such case a conservative rule was adopted: rather than making an unsupported inference it was put into the category "position unknown". Doubtless those unknown instances, of which there were 66 in the whole 1994-2006 period (65 in the 1994-2003 interval), or 18.8% (and 24.2% in 1994-2003) of the total number classified in our data and shown in Table 3.2, are spread across the position categories reported by Rhoades, who, we must suppose had full access to the underlying case information.

We might hope to learn something about the way they were distributed, by comparing the yearly totals from Rhoades' Table 19 by position among all cases investigated with the corresponding figures in our dataset *for the years 1994-2003.* Since Rhoades has no "unknown position" cases while we do, his yearly totals for the number of respondents of investigations for every position must be at least as big as ours. In particular, they will be equal if none of the investigations regarding the position went into the summary with missing or insufficient information to identify the position. For example, in 1995 we have 17 instances where we are unable to determine the position upon reading the case summary. Hence, Rhoades counts for the number of investigations in which respondent holding a given position are involved will be often higher than ours (the 17 issues we group into "unknown position" will be distributed across all of some of the categories).

We can imagine gaining some knowledge of the distribution of our unknown cases by subtracting our "totals by position" from Rhoades` corresponding totals. If, say, the difference between the latter total for the number of investigations in which full professors were involved in 1995 and our number for the same count is 0, we will know that none of the 17 case summary we classified as unknown position related to a full professor. On the contrary, if the difference is positive, we will know how prominent is the contribution of full professors to our category "unknown position".

Implementing this procedure, however, proved to be problematic for a number of reasons that make our data and Rhoades (2004) not perfectly comparable. In the first place, it is not clear how Rhoades has classified individuals in the senior positions within the "non-academic category" (as we labeled it) – i.e., project managers, research scientists, etc. Have these been aggregated with "staff", or has it be possible to use unpublished information to precisely classify all the positions? Second, the numbers do not square, and a example taken from 2003 will show: we focus on positions whose definition should not be controversial, and find that Rhoades (2004: Table 19) lists 2 Professors, 2 Associate Professors, 2 Assistant Professors and 4 Postdoc fellows as involved "in misconduct investigations". But our count for individuals in these academic positions on the basis of the published ORI figures for 2003 is: 4 Professors, 5 Associate Professors, 2 Assistant Professors and 3 Postdoctoral Fellows. The

details in the case summary section of the ORI Annual Report for 2003, relating to closed investigations in that year confirms those totals: 1 Professor in investigations with findings (Radolf) and 3 involved in investigations without findings, for a total of 4; 1 Associate Professor involved in investigations with findings (Gelband) and four involved in investigations without findings, for a total of 5; 2 Assistant Professors involved in cases without findings, for a total of 2; 3 Postdocal Fellow involved in cases with findings (Koltover, Rooney, Smith) and none in cases closed without findings, for a total of 3. What appears in the ORI reports and in Rhoades' Table 19 evidently are not the same for 2003, but the reasons for these divergences cannot be found from the publicly available data.

9. Tables 23-26 in Rhoades (2004) are distributions conditional to the highest degree held by the respondent. We did not use this information in our micro-data (it was not always possible to infer it from the case description for investigations closed with findings and almost never mentioned for investigations closed without findings).

10. Tables 27-30 in Rhoades (2004) show distributions of cases by gender. Several points should be noted here. First, from the published case summaries y of investigations closed without findings it is never possible to infer gender, but clearly Rhoades has been able to access and make use of that information. Second, even if we had access to the same information, our figures would come from aggregations of the micro-data for a restricted set of cases – relating to the three forms of scientific misconduct, and there could be small discrepancies due to our exclusion of cases (closed without a misconduct finding) that had involved other charges, e.g., sexual harassment**.**

11. As a check on the quantitative importance of that source of discrepancy, it was possible to count the number of males and females for investigations closed with findings, a piece of information that is available from both **Rhoades' Table 29**, and from our microdata, as we always have been able to determine the gender for respondent in investigations where there was a finding of misconduct, because it is possible either to make an inference from the name or, if that is ambiguous, statements such as "Mr. X has agreed to exclude himself/herself" that are recur throughout these summaries.

Table A2 (below) shows the resulting annual counts. Our over-all total for both genders coincides with the 133 reported by Rhoades (2004) for the 1994-2003 period, but it will be seen that the subtotals by gender are slightly different: Rhoades' table 29 shows 91 males, 41 females, and 1 unknown. One of the two extra males in our total may correspond to Rhoades "unknown" case, which is perplexing in itself, inasmuch as the gender was unambiguous from the summaries. We appear to be 1 female short in our total vis-à-vis Rhoades', and indeed, manually counting the cases from the ORI Annual Reports 1994-2003, the total number of cases for 1994-2003 is 134. What explains this?

Table A2: Annual Counts of male and female respondents in ORI investigations closed with findings of misconduct, as compiled from the Pozzi-David (2007) dataset

Year	Males	Females
1994	9	2
1995	15	10
1996	11	6
1997	12	2
1998	4	5
1999	6	6
2000	6	1
2001	10	3
2002	11	2
2003	9	3
Total	93	40

The source of the difference lies in the double report of the ORI's findings regarding Shaan F. Munjee, a female research fellow against whom allegations of fabrication and falsification were substantiated. The case is described by two separate summaries of findings, that appeared in the Annual Reports for 2001 and 2002. The texts of these reports are identical, and our compilation viewed this as a publishing error; consequently this case was entered only once (for 2001) in our micro-dataset.

12. Tables 31-32 in Rhoades (2004) tabulates the administrative actions taken, and in **Tables 33-44** the analysis that was carried out for respondents is repeated for those who were the "whistleblowers" in the cases investigated. The available case data from the published summaries precluded our analysis of these questions at the micro-level.

13. Tables 45-48 and **Tables 53-56** deal with the length of the process for inquiries and investigations, respectively, whereas in **Tables 49-52 and 57-60,** Rhoades presents statistics on the sizes of the ORI panels for inquiries and investigations, respectively. This was material that is not available in the summary reports on individual cases, and therefore is not in our dataset.

Conclusions

There are only three sets of tables in our paper that overlap directly with those in Rhoades (2004) study, and for which it might be supposed that the statistics shown would coincide perfectly. That, however, is not always a warranted supposition.

The set of Rhoades' Tables 1-5 is indeed one with which our data should be fully consistence, as the aggregate time-series measures are defined in the same way and the same ORI sources appear to have been consulted. But whereas the figures agree perfectly in many instances, in

other they do not. Sampling the divergent instances at random, it has been shown that one generally can find the source of the Pozzi-David totals in the ORI Annual Reports for the year in question, or, and a small number of cases, in the variant figure that is cited for the date in question in a subsequent Annual Report. We have not been able to do this for Rhoades' totals in those cases of discrepancy; if his study has drawn upon revised numbers (to which we have not had access), no mention of the existence of post-publication official revisions has been found in the ORI Annual Reports for 1994-2006. Further, in the case of one quantitatively striking discrepancy in the number of allegations made to the ORI in 2003, the published total count that is reproduced in Table 1 of Rhoades (2004) is traceable to a mistake in the ORI Annual Report itself, which is corrected in our dataset.

In the cases of comparisons between the aggregates reported on the basis of our dataset and those in Rhoades Tables 7-8, and 19-22, the observed discrepancies stem from a divergence between the variables whose distributions are being reported: in Rhoades' tables what appears are the frequencies of cases that involved specific misconduct charges, whereas in Pozzi-David (as the Notes and Sources to Graph 2.1 and 2.2 make clear) it is the frequencies of the specific charges that were involved in the set of investigations closed in the period. Thus, where Rhoades (2004) would treat a case involving fabrication and falsification as one instance of that (combined) form of misconduct, frequency counts obtained for that period from the Pozzi-David micro-level dataset treat count two distinct charges in that case, one instance of falsification and one of fabrication.

Even though this independent study has not had the benefit of the knowledge and access to data that and ORI staff researcher would enjoy, it has been shown to be possible to use the published ORI statistical reports in a consistent and accurate fashion to extract findings that hitherto have not been forthcoming from analyses of that data. It goes without saying that providing qualified and independent social scientists with more extensive access to the underlying data, with due precautions for the privacy of information from investigations that were closed without findings of misconduct, would contribute to improving public understanding of the regulatory work of the ORI and illuminate the phenomenon that this agency is seeking to address through its other, educational and training activities.

Appendix B

On the Durations and Potential Resources Costs of ORI Inquiries and Investigations into Allegations of Scientific Misconduct, 1994-2003

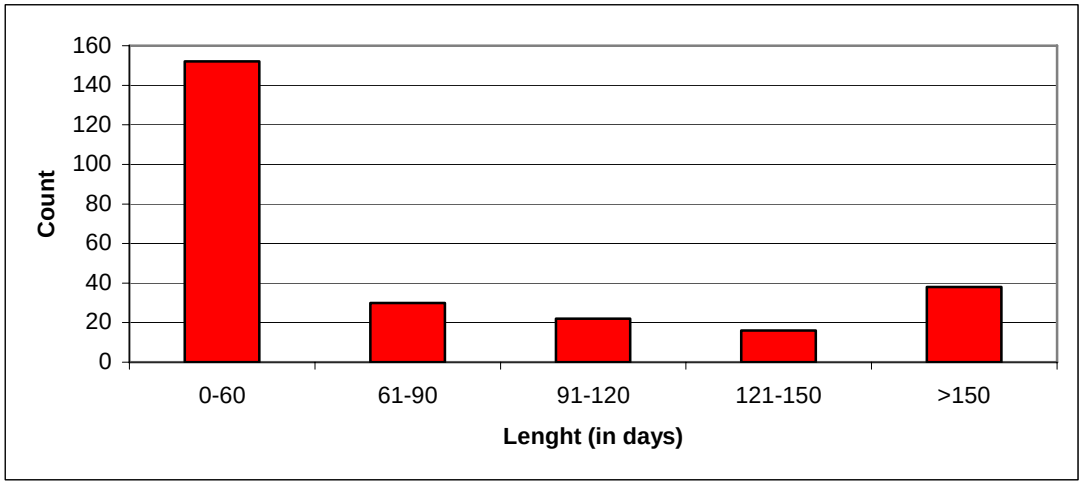
The statistical review provided in Rhoades (2004) of the ORI's handling of cases of alleged scientific misconduct in the period 1994-2003 presents annual tabulations of the durations of the inquiries that were opened following the Office's receipt of allegations. Not all reports of misconduct led to the opening of an inquiry, and not all inquiries turned into investigations. Rhoades, in Tables 45-48 provides annual counts of the numbers of cases by duration intervals for all inquiries, for inquiries that led to investigations, and also for investigations that led to findings of misconduct. For each of the three kinds of "inquiries" we have aggregated these counts across the whole period to obtain distributions of cases by duration intervals. As there are indications of some shifts between the distributions on either side of 1998, we also calculated the distributions for the period 1999-2003 -- on the ground that these were likely to be more relevant to the latter part of the 1994-2006 time-span covered by our dataset.

There are two quite different issues on which our study touches that this data can help illuminate. One is concerned with our interest to learn what can be discovered from the publicly available ORI data about the length of the detection and correction lags in cases of proved scientific misconduct, where, by "correction" we mean a formal public notice that indicates that the archival record is misleading as a result of a deliberate breach of scientific research norms. Such information has at least some bearing on the question of the magnitude of the social costs that ensue when other researchers are misled by such false statements in the published research record. The second issue relates the social costs of the process of determining whether allegations of misconduct that would have left false and misleading statements in the research record can be substantiated. Learning about the duration of that process, and the resources that are occupied with it is at least a first step toward gauging those costs for the representative case of misconduct.

1. Estimating the Duration of Inquiries in Cases of Substantiated Misconduct

We would like to estimate the mean and median duration of the period between the opening of an inquiry and the closing of the investigation to which that inquiry led, when the investigation closed with findings of misconduct. That duration would be what we have suggested should be subtracted from the "correction lag" – which we have defined for purposes of measurement from the available data as the time elapsed between the detection of the misconduct in a (journal) publications and the public "correction" of the misconduct by either the ORI's announced findings, or a possible prior voluntary "retraction" on the part of the respondent in the case being investigated. Because our estimates of the "correction lag" are constructed from the data pertaining to cases in which investigations were closed with findings of misconduct, information on the distribution of durations of the inquiries that ended in that manner is what should be examined in this connection.

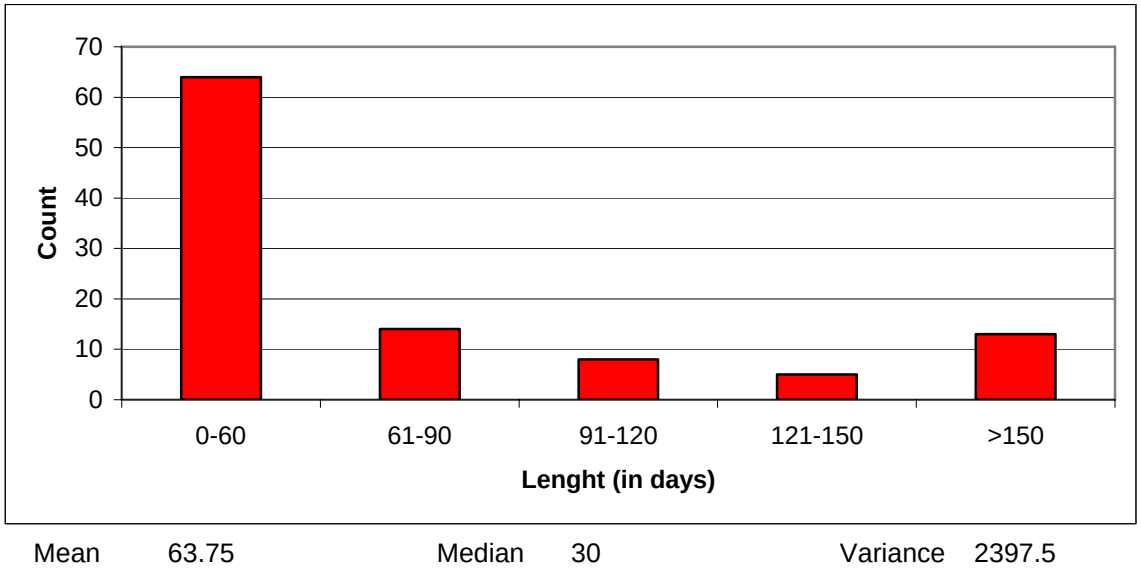
Of course, inasmuch as the data provided by Rhoades (2004: Table 46-48) on that point pertains only to the duration of the inquiry phase, it is likely to seriously understate the duration of the inquiry and investigation. The following Figure shows the distribution for all Inquiries, from Rhoades (2004: Table: 45), where the pronounced mode in the 1-60 interval presumably reflects the many cases where the inquiry was quickly terminated because there was strong *prima facie* indications of misconduct --warranting initiation of an investigation without further delay.



Mean 68.02 Median 30 Variance 2647.6

Graph B1: Distribution of the length of inquiries that resulted in investigations, 1994-2003
Source: Elaboration of Rhoades (2004): Table 46
Note: Mean, median and variance of the distribution are computed assuming that the actual length falls at the midpoint of the interval. For example, the length of all the investigations in the 0-60 days interval is assumed to be 30 days. In order to compute those statistics, we also capped the "more than 150 days" conservatively at 180 days.

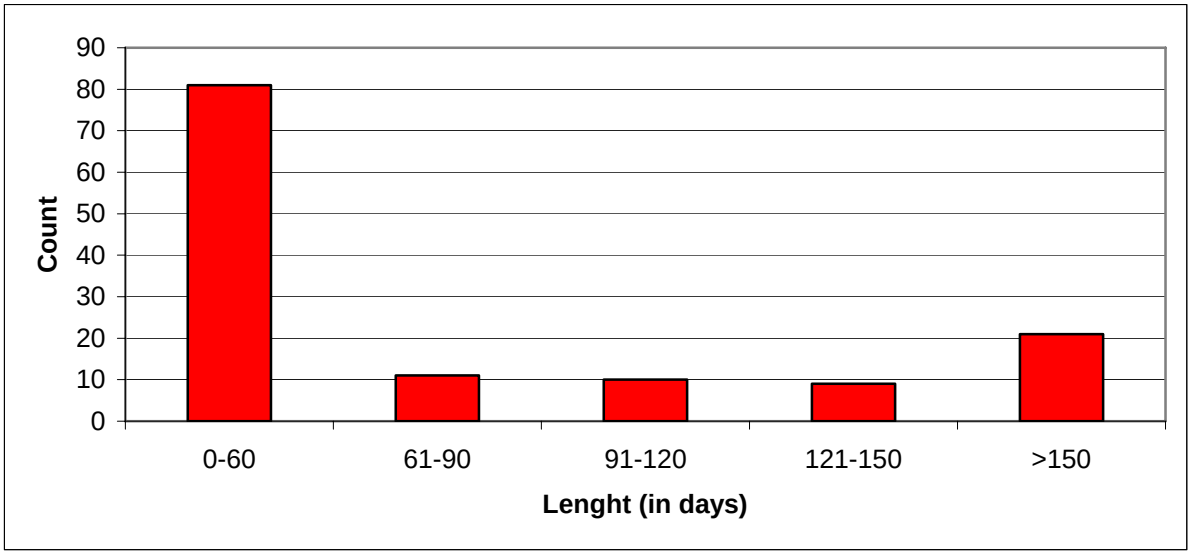
As will be seen from the following distribution data for the post 1998 period covered by Rhoades' study, there was a slight downward shift in the durations distribution over time, and if we wish to allow for this in fixing mean estimates appropriate to the whole period up to 2006, the later sub-period estimates would seem the appropriate ones to use.



Graph B2: Distribution of the length of inquiries that resulted in investigations, 1999-2003.
Source: Elaboration of Rhoades (2004), Table 46.
Note: Mean, median and variance of the distribution are computed assuming that the actual length falls midway through the interval. For example, the length of all the investigations in the 0-60 days interval is assumed to be 30 days. In order to compute those statistics, we also had to cap the "more than 150 days"; the upper limit was set at 180 days.

As a consequence of simply using the inquiries duration approximate the typical duration of the inquiry and investigation in the representative case where misconduct findings were returned by the investigating panel, the adjusted "correction lag" will remain an over-estimate. Indeed, it is likely to be a substantial over-estimate of the detection lag in those cases where the opening of the inquiry followed shortly after the report of an allegation that coincided with the journal publication. On the other hand, the extent of that over-estimate would be reduced where the reported allegation lagged the journal article's publication, and when the latter lag was protracted, it is quite possible that the adjusted "correction lag" would under-state the detection lag. The plausibility of that situation cannot be established a priori, so we need to look at the magnitude of the implied lags in order to venture a judgment on this point.

By comparing the foregoing with the distribution of all the inquiries *that led to investigations*, it should be possible to see whether there was a tendency for the inquiry phase to be shorter when the evidence is strongly pointing to substantiation of the allegations—i.e., they move them into investigations. Focusing on inquiries that led to investigations removes the inquiries that are quickly dismissed without investigations, thereby exposing the effects of the other considerations that similarly may work to truncate the inquiry phase.



Mean 68.07 Median 30 Variance 2818.18

Graph B3: Distribution of length of inquiries resulting in investigations closed with findings, 1994-2003.

Source: Elaboration of Rhoades (2004), Table 47.

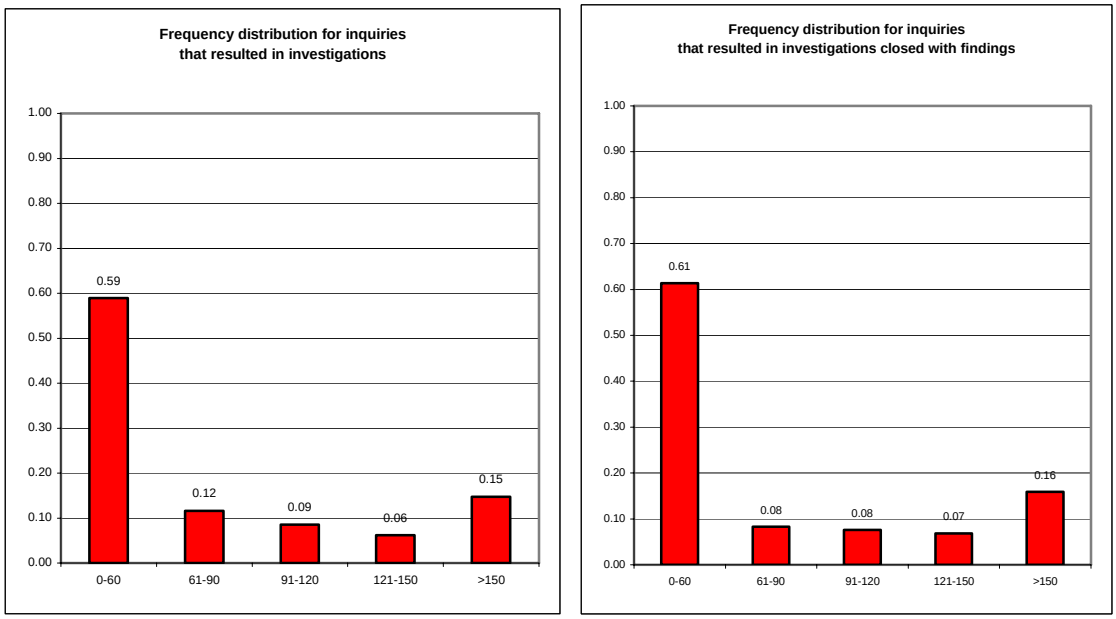
Note: Mean, median and variance of the distribution are computed assuming that the actual length falls midway through the interval. For example, the length of all the investigations in the 0-60 days interval is assumed to be 30 days. In order to compute those statistics, we also had to cap the "more than 150 days"; the upper limit was set at 180 days.

As can be seen, the proportion of inquiries within the shortest interval (0-60 days) is slightly higher for the inquiries that led to investigations closed with findings, but the mean duration length for this group is 68 days, compared to 63 for all inquiries that continued as investigations, is also higher, suggesting that there are more inquiries in the right tail. Comparison of the frequency distributions for all the inquiries that led to an investigation and for those who led to investigations closed with findings just remarks the striking similarity between the two groups.

It will be recalled that the means for inquiries underestimate the magnitude that we are seeking, and that it is therefore necessary to estimate the mean "investigation period", so that it may be added it to the corresponding mean duration of inquiries, and use that to correct the "corrections lag" estimates.[18] Rhoades (2004), in commenting on the data in his Table 52,

[18] Adding these estimates of the means will yield an estimate of the mean duration of the entire inquiry and investigation period (when an investigation was conducted) only if the two distributions were independent (no covariance between the two sub-intervals). Unfortunately that cannot be established on the basis of the available data. If there is positive covariance, the procedure of adding the two means will understate the mean of the combined period, the opposite bias would result where there is negative covariance. *A priori* considerations would seem to favour positive covariance, in that cases where the prime facie evidence was strong would tend to move more quickly into to the investigation stage and then be more rapidly closed (with a finding of misconduct), whereas less clear-cut cases would occasion more lengthy deliberations in each stage. This suggests that the procedure adopted would yield estimates that, on being subtracted from the "correction lag" would tend to impart a downward bias to the residual – which is our estimate of the mean duration of the "detection lag".

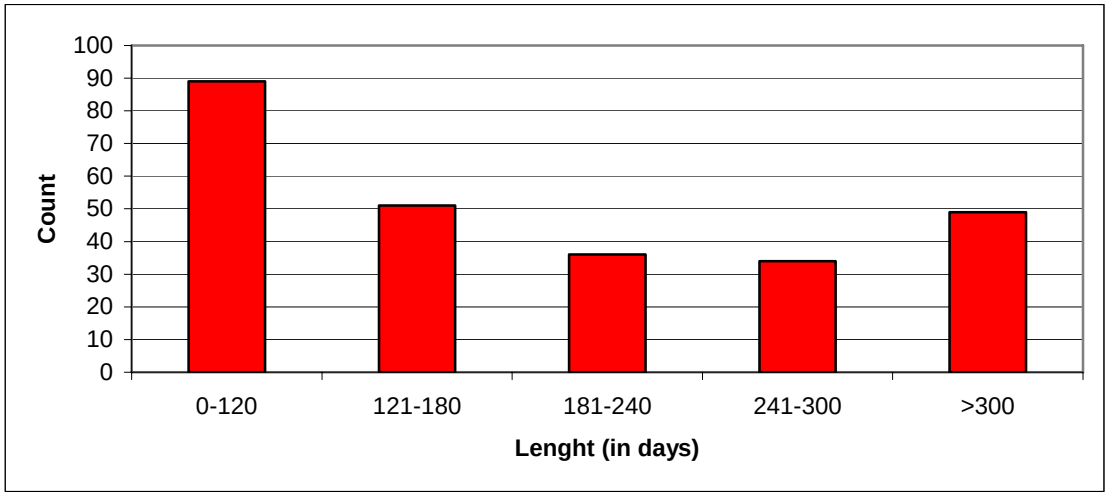
notes that while the PHS regulation for investigations of misconduct say that "an investigation should ordinarily be completed within 120 days of its initiation" (which he takes to be the date the investigating committee's first meeting), the proportion of all the investigations conducted by the ORI during the 1994-2003 period that actually conformed to this notional 120-day standard (0.34) was much the same as the proportion (0.32) that took more than twice as long to be closed. This may be seen from the frequency distributions displayed in Graph B5.



Graph B4: Comparison of the frequency distribution of the length of inquiries, 1994-2003
Note: Shown on the left for all the inquiries resulting in investigations, and on the right, for inquiries closed with findings of misconduct
Source: Elaboration of Rhoades (2004), Tables 46 and 47.

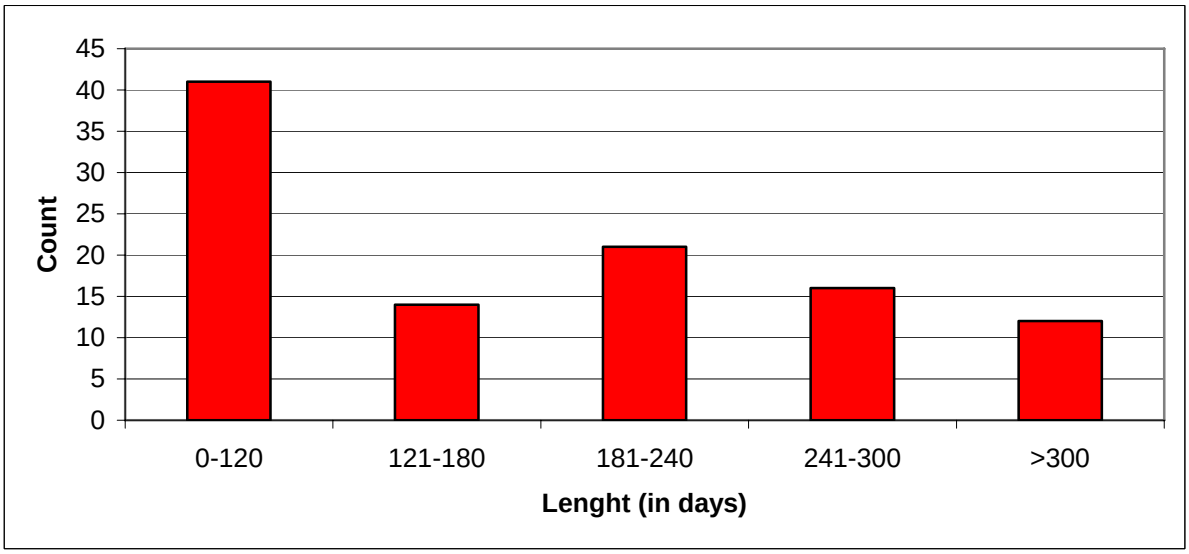
The corresponding distribution plot for the years 1999-2003, shown in Graph B5, indicates that the relative importance of those protracted investigations diminished after the 1990s, bringing down the mean for all investigations. The distribution of durations for investigations that were closed with findings of misconduct is the one that is most immediately relevant for our present purposes, and we wish to combine its mean with the counterpart for inquiries that led to investigation, and compare it will the estimated correction lag in the case of investigations that closed with findings of misconduct.

It is seen from the following that frequency bar-diagram that for the latter the mean investigation period is shorter than that for all investigations, and although it shows similar bimodality the proportions of cases that fell within the 120 day standard and those that more over twice that in length, the upper and lower modes are not so symmetric: very protracted cases were not as frequent as those that were dispatched comparatively quickly.



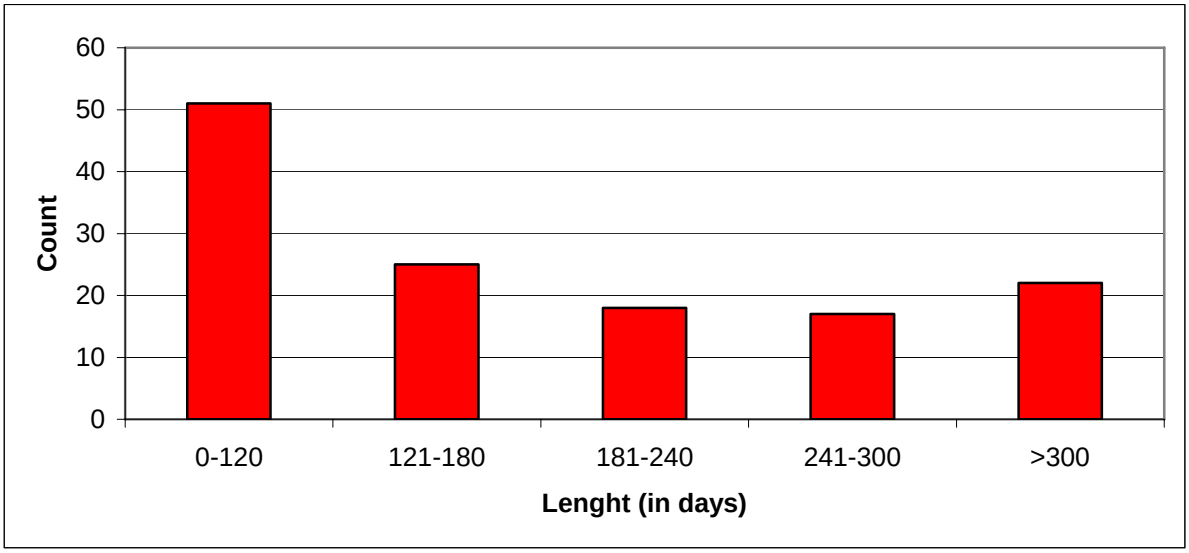
Mean 177.22 Median 150 Variance 10603.87

Graph B5: Distribution of the length of investigations, 1994-2003.
Source: Elaboration of Rhoades (2004), Table 54.
Note: Mean, median and variance of the distribution are computed assuming that the actual length falls midway through the interval. For example, the length of all the investigations in the 0-120 days interval is assumed to be 60 days. In order to compute those statistics, we also had to cap the "more than 300 days"; the upper limit was set at 360 days.



Mean 165.87 Median 150 Variance 9715.75

Graph B6: Distribution of the length of investigations closed with findings, 1999-2003.
Source: Elaboration of Rhoades (2004), Table 54.
Note: Mean, median and variance of the distribution are computed assuming that the actual length falls midway through the interval. For example, the length of all the investigations in the 0-120 days interval is assumed to be 60 days. In order to compute those statistics, we also had to cap the "more than 300 days"; the upper limit was set at 360 days.

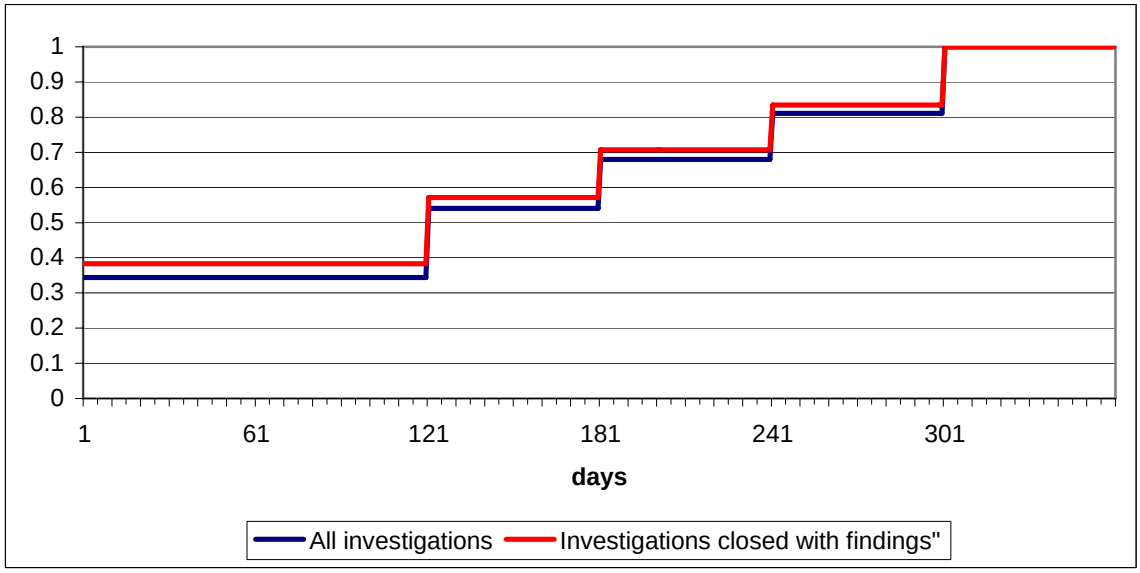


Mean 168.72 Median 150 Variance 10521.84

Graph B7: Distribution of the length of investigations closed with findings, 1994-2003.
Source: Elaboration of Rhoades (2004), Table 55.
Note: Mean, median and variance of the distribution are computed assuming that the actual length falls midway through the interval. For example, the length of all the investigations in the 0-120 days interval is assumed to be 60 days. In order to compute those statistics, we also had to cap the "more than 300 days"; the upper limit was set at 360 days.

A comparison between the cumulative distribution functions (in Graph B8, below) for the full set of investigations and the proper subset closed with finding makes it apparent that the process of arriving at positive findings of misconduct is unambiguously faster than that of closing investigations without findings, since there is first order stochastic dominance of the distribution for investigations closed with findings.

Unfortunately, it is not possible to determine whether this tendency became even more pronounced in after the 1990s, as was found for all investigations, because the corresponding data for the 1999-2003 interval is not available in Rhoades' (2004) tables. We might suppose, however, that the mean for the 1994-2003 period in this case would tend towards overstatement of the average duration of investigations that closed with findings during the years 2000-2006.



Graph B8: Cumulative density step-function for the length of investigation intervals, 1994-2003
Source: Elaboration of Rhoades (2004), Table 54 and 55.

The upshot of the foregoing calculations is that the average length of the ORI "inquiry and investigation" process for cases that end up with findings of misconduct is (68.07 + 168.72=) 236.79 days, or 0.65 of a year. Consider, then, our estimate of the mean "correction lag" -- measured as the time elapsed between the date of the anterior journal publication and the date of a voluntary retraction or, in the latter's absence, the close of the ORI investigation with findings of misconduct, involved in case – which was found to be 3.65 years for the period 1994-2006. Subtracting the adjustment for the inquiry and investigation process therefore yields 3 years as the estimated mean "detection lag."

As reasons have been given to supposing that the downward adjustment may too large, on account of positive covariance in the durations of the inquiry and investigation phases, and also due to the indications of a trend toward shorter average investigation durations in the post 2000 period, this would suggest that 3 years is perhaps somewhat too low an estimate. That view, however, is conditional on accepting the opening of an inquiry by the ORI as the close of the detection lag measured from the date of publication. But inasmuch as preliminary institutional inquiries will have preceded reporting of allegations to the ORI, on this account it could be held that the 3-year estimate may not be too low, and, indeed might well be excessive. We therefore propose the "Optimistic Goldilocks' Solution": "It could be just about right."

2. Estimating the Panelist Members' Time per Case in ORI-Supervised Inquiries into Reported Allegations of Scientific Misconduct

An opportunity to try to gauge the order of magnitude of the direct resource costs borne by institutions that report and inquire into allegations of misconduct under ORI supervision is afforded by Rhoades (2004) study, which has compiled and made public some of the necessary systematic data about the sizes of the panels involved. This information pertains to

(i) inquiries that led to an investigations (Table 50, split into the two sub-samples 1994-1998 and 1999-2003), (ii) inquiries that led to investigations closed with findings (Table 51, only for the whole 1994-2003 interval), (iii) all the investigations (Table 58, split into the two sub-samples 1994-1998 and 1999-2003), and (iv) all the investigations closed with findings (Table 59, only for the whole 1994-2003 interval). In the following we detail our exploitation this data to compute the average size of the panel for both the inquiry and the investigation step, and combine these separate figures into an estimate the universe for of those cases the cases ultimately yielding findings of misconduct. The results of the latter calculations are shown below in Table B1.

	1994-1998	**1999-2003**	**1994-2003**
Inquiries that led to an investigation	2.82	3.14	2.95
Inquiries that led to an investigation closed with findings	na	na	2.54
All investigations	3.56	3.78	3.65
Investigation closed with findings	na	na	3.29

Table B1: Average size of the panel for inquiries and investigations.
Note: In order to compute the averages, we had to cap the interval "six panelists or more". We assumed that the average size of a panel in that group is 6.5 panelists.

For the whole sample period 1994-2003 the mean was 2.95 panelists per investigation, whereas the two sub-periods average 2.82 panelists, and 3.14 panelists, respectively. This change reflected a rise in the share of investigations that involved panels of 5 and more, and it is interesting to note that this rise in panel size was concentrated among the investigations that eventually were closed *without* findings. If that trend continued, it would seem more appropriate to work with the higher point estimate of persons per investigation, which we can put at 3.65 persons.

It is interesting to consider what the estimates of the average durations of these investigations (discussed in Section 1, above) might imply about the resource inputs that are entailed in the process for the average misconduct inquiry -- when they are brought together with Table B1's

estimates of the mean number of panelists involved. To venture to produce more that heuristic estimates on this question is surely hazardous if one does not know with what frequencies the panels are convened during the investigation phase. Such data in not published by the ORI and, indeed, may not be collected by any agency, because the investigations typically are conducted by the relevant institutions under ORI supervision. Systemic collection of representative data would seem to require gathering the records of stratified samples of panel meetings from the various institutions that are involved – a formidable task.

Therefore, as hazardous as it may be, it is tempting to ask what would be the number of panel-person days devoted to an average investigation that closed with findings of misconduct – if one supposes that the panel's member were convened, on average, for as much as one 6-hour day in each month of the time span that that the case was under examination. From the data exhibited in Graphs B4 and B5 of the preceding section it was concluded that the inquiry and investigation process in such cases averaged 236 days in duration, suggesting an average of 7.9 meetings, or 47.2 hours of panel sessions per case under our assumption.

Putting that figure together with our estimate of the average panel size, we arrive at the figure of 172 of panelist hours per case, implying that the 165 ORI investigations that closed with findings of misconduct during the years 1994-2006 would have occupied 28,472 hours of panelists' time.

Having gone this far, we can go farther, by finding the corresponding "guesstimate" for the investigations that were closed without misconduct findings -- which numbered 102 *in toto* during the same period. Since we have found (from the data in Graph B3) that mean duration of all investigation was 177.2 days, by considering at as the weighted average duration of the investigations closed with findings (168.7 days) and those closed without findings, we can infer that the mean for the latter group was 191 days during the period 1994-2003. We might accept that as applicable for present purposes, and add to it the estimate of a 68.02 inquiry period for cases that in which an investigation ensured, giving a 259 day mean total duration for the average "panel life." That translates into 51.8 hours of panel session per case, under the same assumptions as were previously applied. When combined with mean estimate of 3.78 persons per panel shown for all investigations in Table B1,[19] this yields 195.8 person hours per case, implying that the 102a ORI investigations that closed *without* findings of misconduct during the years 1994-2006 would have occupied 19,972 hours of institution-based panelists' time.

Adding the two components, the 267 ORI closed investigations during 1994-2006 occupied 48,444 hours of panelists' time. Supposing, for the sake of argument, that a salaried academic work-year is reckoned as having a 9 months duration, or 184 (= 245x 0.75) 8-hour workdays, we can translate that total person-hour input as 32.9 salaried academic work-years. This implies that 8 closed investigations consume the equivalent of a year of salaried academic work-time.

---

[19] The rationale for this is that, as has previously been noted, in those cases were the investigations terminated with misconduct findings the durations were tending to lengthen during the 1994-2006 period, as well as being generally involving larger panels than the investigations that found misconduct.

That figure, however, ignores the face that not all opened inquiries proceed to the institutional investigation stage, and one would suppose that inquiries that do not become investigations are, on average, disposed of more quickly than the ones that do. Indeed, this appears to be true. From Rhoades (2004: Table 1), it may be inferred that 70 inquiries conducted by the ORI in the period 1994-2003 that did not lead to investigations. We have no direct estimates for the mean duration of these inquiries, but can infer from the average durations of all inquiries, and those that did lead to investigations that the average length of the inquiry phase for cases of this kind was 59.93 days, which we reckon as 12 session hours. The (similarly inferred) mean size of these panels is found to be 2.96 persons,[20] so that for the 70 cases that were disposed of by this initial component of the ORI inquiry process we arrive at a total panelist-time requirement of 2483 hours, or 1.70 salaried academic work-years.

So, all in all, the 337 cases that ORI opened for inquiry and disposed of during the 1994-2006 period would, on the foregoing speculative assumptions, have consumed some 34.6 (=32.9 + 1.7) salaried academic work-years of the attention of senior scientific personal just for the investigative processes. Is one such person-year for each ten cases processed a big number, or a small one?

Viewed from one perspective this figure clearly might be suspected of over-stating the time requirements: we simply calculated the number of panel sessions per case, and multiplied by the total number of cases to obtain an aggregate estimate of the panel session hours. Our calculations therefore did not (and, given the data constraints could not) make any allowance for the economies of processing cases in parallel and dispatching them in batches at panel meetings. This was based on the understanding that each investigation, at least, would have been dealt with by its own panel. Might there have been instances in which a given panel handled multiple cases concurrently? Where that a general practice, a proper adjustment for the deliberations on more than one case in a single panel session conceivably reduce the total time requirement by as much as one-half – if more than two cases were handled concurrently in some fraction of the cases, and the entire session was devoted to one case on the other occasions.

But, from another vantage point, the illustrative time requirements presented here should be seen to be too narrow in scope, and hence tending toward being excessively modest in magnitude. They omit the time of witnesses in these cases, not to mention that of the respondents, and that of the administrators of the respondents' institutions. Moreover, they suppose that the only time input required of the panelists is that of attending the meetings, rather than making an allowance for the time devoted to reading the documents in preparation for deliberations when the panel convened. The latter omission may be particularly serious, and rather resembles the mistake of non-academics in reckoning the work time of teachings only in terms of their classroom hours. Suppose that we had understated the time actually required by 50 percent, then, the two major sources of potential bias in the illustrative estimate would turn out to have just offset one another.

[20] These numbers come from weighted averages of the figures for the 259 inquiries that led to investigations and the 70 inquiries that did not lead to investigation.

Ignorance does not entitle us to claim that such fortuitous outcomes can be expected, so we cannot say whether the truth is exaggerated or unstated by an average figure of one academic work-year being required to carefully scrutinize each batch of 10 cases of alleged scientific misconduct. Our intention in producing an illustrative estimate of that kind has been, to be sure, the ostensible purpose of inviting other to think about what might be plausible orders of magnitude for the resource costs of a well structured and professionally conducted process such as that which has been established by the ORI.

But, the deeper and more serious purpose has been to indicate what could be learned by integrating various bits of information from systematic data collection, and supplementing it with estimates of particular parameters derived from small scale surveys to determine the actual use of the time of research scientists and administrators the array of entailed activities: attending panel meeting conducted by the grant-receiving institutions, studying the relevant documents and testimony in cases where allegations merited formal inquiry, and preparing appropriate reports for submission to the PHS through the ORI.

Were quantitatively research efforts to stop at the point where the preceding exercise halted, one would be left with what can be seen as estimates of the "human time-input requirements" -- a an aspect of the "process engineering requirements" for inquiries into allegations of scientific misconduct. To move beyond that and gauge the economic resource costs, it would be necessary to know the detailed composition of the institutional panels and the respective "opportunity costs" of the mix of expertise that these proceedings typically engage – that is, the incremental valuation of the participants' time that otherwise could be devoted to research, teaching and administrative pursuits in their respective institutions. Undoubtedly these measures of social resource costs will be far from uniform across fields and scientific domains. Consequently, although it would be familiar for economists and feasible in principle to pursue the approach of using relative academic salary rates to weight the contributions of panel members of different rank, experience and familiarity with the research methodologies in cases of alleged fabrication and falsification, in practice this would constitute a non-trivial research exercise.

A Conclusion

Are social science research inquiries of this nature really called for? The question is perhaps best answered by posing another one: If a representative opportunity cost figure for a senior research scientist's "academic work-year" were put at $150,000, and the institutional burden of conducting careful inquiries into allegations of research misconduct in the biomedical and behavioral sciences in the U.S. therefore might be in the neighborhood of $15,000 *per case*, would it be such a good idea to proceed to expand the conduct of such proceeding on a global scale without first trying to establish what the direct costs of such a policy actually might be?

Appendix C
On the gender and the whistleblowing activity

This appendix deals with two issues that we planned to take analyze to some extent but were unable to do, due to the lack of suitable data.

C.1. Whistleblowing

The activity of whistleblowing is of sure relevance to the topic of misconduct; it is, after all, the moment when everything begins. In the microdata, there is no mention of the identity of the whistleblower; hence, our only shot at the issue relies on tabulations reported by Rhoades (2004) in table 35. Rhoades shows the distribution of whistleblowers by position and distinguishes the allegations that were in the end substantiates and from those that were not. On this basis, he is able to build a "substantiation rate" and comment on differences in this rate across positions. Are some categories really blowing the whistle more inaccurately than others? Our calculations, based on Rhoades own data, displayed in Table C1 suggest caution in interpreting differences in the substantiation rate across groups. Even though the numbers seem indeed to be different, the differences are not statistically significant. There is actually very little we can say about the giving a point estimate to the substantiation rate for a given category.

C.2. Gender

It is interesting to look at the distribution of investigations by gender in our microdata. Since we have no variation in the outcome of the investigation for cases where the gender is known[21], we will not be able to use gender as a control. Nevertheless, it can be insightful to take a look at the distribution of findings of misconduct across gender, as we do in Table C2. The problems in commenting such a table arise from the fact that there are a number of stratified selection processes behind those numbers. The nearly absence of female performing misconduct at the higher level is most likely due to the under representation on women in that group, rather than with other aspect truly related with misconduct. In the same fashion, the prominence of women in the share of crime for the category "staff" is linked to the fact that we are looking at biomedical sciences and women are widely represented in the category through nurses. Perhaps positions as graduate students and post-doc are less prone to the concerns of gender induced occupational segregation. Looking at those two rows, no gender driven specification seem to emerge: women perform more falsification, just as their male colleagues.

We would have been interested in combining the two topics of this appendix, looking at whether significant differences can be found in the presence of female among respondents and whistleblowers. This proved to be not possible; in fact we have no way to match tabulation in Rhoades about distribution of respondents by gender with distribution of

[21] Actually gender is never explicitly mentioned even in summaries for investigations closed with findings. However, in that cases we were able to infer it either from the name of the respondent or thanks to expression such as "…agreed to exclude himself (herself)…".

whistleblowers by gender. Table 28 in Rhoades provides tabulation of gender of the respondents in misconduct investigation, with breakdown into the two periods 1994-1998 and 1999-2003. Table 42 in Rhoades (2004) provides an analogous tabulation but regarding the gender of the whistleblower in misconduct investigations. Our aim had been to compute the ratio of whistleblowers to respondent by gender and check whether it changed in the two subperiods. Two things should be noted, however:

1) The total number of whistleblowers adds up to 178 in the 1994-1998 and to 111 in 1999-2003. The same figure for the number of respondents adds up to 170 and 104 respectively. This could be explained if investigations have multiple whistleblowers more often than they have multiple respondents or if the distribution of the number of whistleblowers has a right tail thicker than the distribution of the number of respondents.
2) The gender of the respondent is known in every case but once per subperiod while the gender of the whistleblower is unknown 38 times in 1994-1998 and 23 times in 1999-2003. This is not surprising, since we can imagine that protecting the identity of the whistleblower is an important issue for those who perform the investigation.

We want to consider only the cases where the gender of the whistleblower is known. Hence, we would want to discard investigations where the gender of the respondent is known but the gender of the whistleblower is not. Only, we do not know which ones those cases were, since we have only aggregate numbers and not descriptions of specific cases, in particular we ignore whether they involved males or females as respondents. This makes it impossible for us to advance any further on the issue. Researchers with better access to the data might be able to do more and provide some evidence on this potentially relevant question.

Institutional Position of the "Whistleblower" in the case that was investigated	Number of whistleblowers whose allegations were substantiated	Total number of whistleblowers	Mean probability of substantiation	Standard dev. of the probability of substantiation
All Cases (unconditional on position)	108	216	0.500	0.497
Faculty appointment: all	85	165	0.515	0.500
Tenure faculty ranks	73	137	0.533	0.499
Non faculty researcher: all	23	51	0.451	0.498
Research Assoc/Assist & Students	14	23	0.609	0.488
Post doctoral Fellows & Technicians	9	28	0.321	0.467

Table C1: Frequency and mean probability of ORI investigation' substantiation of allegations of scientific misconduct during 1994-2003: conditional on the research position of the allegation's source (i.e.,the "whistleblower's" position)

Source: Elaboration of Rhoades (2004), Table 35.

Note: A two tail t-test of diff based on pooled (unequal) sample variance has been performed to test significance in the difference of the means between the group "Faculty appointment: all" and "Non faculty researchers: all"; the means are not statistically significant (P- value from the test is 0.4254). An analogous test has been performed for the two groups "Research Associate/Assistant & Students" and "Post doctoral fellows & Technicians"; in this case the hypothesis of equality is also rejected (P-value 0.2351).

		1994-1999				2000-2006		
		Falsification	Fabrication	Plagiarism		Falsification	Fabrication	Plagiarism
Professor (Full)	Total	6	3	0		6	4	1
	(M,F)	(6,0)	(2,1)	(0,0)		(6,0)	(4,0)	(1,0)
Associate professor	Total	4	3	3		5	2	0
	(M,F)	(4,0)	(3,0)	(3,0)		(5,0)	(2,0)	(0,0)
Assistant professor	Total	2	1	1		3	1	3
	(M,F)	(1,1)	(1,0)	(1,0)		(3,0)	(1,0)	(3,0)
Post-Doc	Total	5	5	1		15	10	2
	(M,F)	(3,2)	(4,1)	(1,0)		(10,5)	(7,3)	(2,0)
Graduate student	Total	6	7	0		7	5	0
	(M,F)	(4,2)	(6,1)	(0,0)		(4,3)	(2,3)	(0,0)
Research assistant	Total	2	2	0		5	5	0
	(M,F)	(1,1)	(1,1)	(0,0)		(4,1)	(3,2)	(0,0)
Undergraduate student	Total	0	0	0		0	1	0
	(M,F)	(0,0)	(0,0)	(0,0)		(0,0)	(1,0)	(0,0)
Research scientist	Total	17	11	2		5	6	2
	(M,F)	(13,4)	(5,6)	(2,0)		(1,4)	(1,5)	(2,0)
Staff	Total	17	10	0		7	7	1
	(M,F)	(6,11)	(4,6)	(0,0)		(7,0)	(2,5)	(1,0)
Unspecified	Total	7	4	0		1	1	1
	(M,F)	(4,3)	(4,0)	(0,0)		(1,0)	(0,1)	(1,0)

Table C2: Distribution of investigations closed with findings, by gender. 1994-2007.
Source: ORI annual reports, 1994-2003